Spring 2015

Low income, supermarket accessibility, and the transportation network: A multimodal analysis identifying areas of poor accessibility and intervention strategies in Indianapolis, Indiana

Andrea Leigh Bailey
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Andrea Leigh Bailey

Entitled
Low Income, Supermarket Accessibility, and the Transportation Network: A Multimodal Analysis Identifying Areas of Poor Accessibility and Intervention Strategies in Indianapolis, Indiana

For the degree of Master of Science in Civil Engineering

Is approved by the final examining committee:

Dr. Konstantina Gkritza
Chair

Dr. Jon D. Fricker

Dr. Jane Frankenberger

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Konstantina Gkritza

Approved by: Dulcy M. Abraham 4/16/2015
Head of the Departmental Graduate Program Date

LOW INCOME, SUPERMARKET ACCESSIBILITY, AND THE TRANSPORTATION NETWORK: A MULTIMODAL ANALYSIS IDENTIFYING AREAS OF POOR ACCESSIBILITY AND INTERVENTION STRATEGIES IN INDIANAPOLIS, INDIANA

A Thesis

Submitted to the Faculty

of

Purdue University

by

Andrea Leigh Bailey

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Civil Engineering

May 2015

Purdue University

West Lafayette, Indiana

To my best friend, Mitchell. Thank you for always believing in me and reminding me of the One through which all things are possible.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has helped and supported me throughout my graduate studies. First, I am greatly indebted to my advisor Dr. Konstantina Gkritza, not only for her financial assistance but for her encouragement and guidance as well. Thank you for giving me the freedom to explore ideas about which I felt passionately.

I would also like to thank the other members of my committee, Dr. Jon Fricker and Dr. Jane Frankenberger. I truly appreciate your time, ideas, and resources that made this research possible.

Finally, I would like to thank my parents, Jeff and Barb Bailey, and my sister Lauren. Thank you for making education a priority, for encouraging me to pursue my dreams, and for all of your moral support.

TABLE OF CONTENTS

Page

LIST OF TABLES ... vi

LIST OF FIGURES ... vii

ABSTRACT ... ix

CHAPTER 1. INTRODUCTION ... 1

1.1 Research Motivation ... 1
1.2 Current Issues ... 3
1.2.1 Obesity and Nutrition ... 3
1.2.2 Vehicles and Transit ... 5
1.3 Research Objectives ... 7
1.4 Research Benefits ... 7
1.5 Thesis Organization ... 8

CHAPTER 2. LITERATURE REVIEW ... 9

2.1 Spatial Analysis ... 9
2.2 Activity-based Analysis ... 12
2.3 Travel Cost Analysis ... 14
2.4 Other Analysis Methods ... 15

CHAPTER 3. DATA DESCRIPTION ... 17

3.1 Geographic Selection ... 17
3.2 Data Sources ... 18
3.2.1 TIGER/Line Files ... 18
3.2.2 IndyGo: Indianapolis' Bus System ... 19
3.2.3 National Household Travel Survey ... 20
3.2.4 Supermarket Locations ... 21
3.2.5 Safety ... 22
3.3 Preliminary Data Analysis ... 24
3.3.1 Transit Network ... 24
3.3.2 Supermarkets ... 24
3.3.3 Crime Locations ... 25

CHAPTER 4. ARCGIS METHODOLOGY AND RESULTS 28

4.1 Background Information 28
4.2 Walking and Driving Analysis 29
4.2.1 Converting the Road Network to a Raster 29
4.2.2 Reclassifying the Road Raster as a Cost Raster... 30
4.2.3 Finding the Cost to the Nearest Supermarket 33
4.2.4 Determining Statistics for Each Census Tract 36
4.3 Transit Analysis... 37
4.3.1 Preliminary Steps 37
4.3.2 Step 1: Distance from the Bus Stop to the Nearest Supermarket 39
4.3.3 Step 2: Distance from Each Bus Stop to the Nearest Supermarket Bus Stop 40
4.3.4 Step 3: Distance from Each Cell to the Nearest Bus Stop 41
4.4 Results 42
4.4.1 Travel Cost.. 42
4.4.2 Travel Cost as a Proportion of Median Household Income........................ 45

CHAPTER 5. STATISTICAL METHODOLOGY AND RESULTS........................ 52

5.1 Exploratory Regression 52
5.2 Ordinary Least Squares 53
5.3 Diagnostics 54
5.3.1 Morans I 54
5.3.2 Normality 54
5.3.3 Heteroscedasticity... 54
5.4 Spatial Regression 55
5.4.1 Weights Matrix 55
5.4.2 Model Specification 56
5.4.3 Lagrange Multipliers... 57
5.5 Model Fit 58
5.6 Marginal Effects 58
5.7 Results 59
5.7.1 Model Significance 59
5.7.2 Marginal Effects.. 62
5.7.3 Implication of Findings... 62

CHAPTER 6. CONCLUSIONS 66

6.1 Summary and Contribution of this Thesis.. 66
6.2 Limitations.. 68
6.3 Recommendations for Future Research.. 69

APPENDIX.. 77

LIST OF TABLES

Table Page

1-1: BMI and its Weight Status Classification 4

3-1: Descriptive Statistics of Selected Socioeconomic Variables 19

3-2: GTFS Reference Files Available for IndyGo, adapted from Google Developers, 2015 20

3-3: NHTS Data Descriptive Statistics 21

4-1: Road Type Classifications (United States Census Bureau, 2015) 30

4-2: Driving Speeds and Costs by Road Type 32

4-3: Descriptive Statistics for Travel Cost by Mode 43

5-1: Descriptive Statistics of Significant Variables 60

5-2: Model Specifications for Each Mode (t-statistics in parentheses) 61

5-3: Marginal Effects of Significant Variables by Mode 62

LIST OF FIGURES

Figure Page

1-1: National, State, and Local Obesity Rates 4

1-2: Commuting Modes of Workers in Marion County Ages 16 and Over 5

3-1: Indianapolis Metropolitan Police Department Districts (IMPD 2010) 23

3-2: Crime Density in Crimes Per Square Mile for Each Census Tract 26

3-3: Crimes Per Square Mile for Each 30m x 30m Cell using a 1-Mile Search Radius 27

4-1: Inset of the Driving Raster by Road Type 32

4-2: Accounting for a Supermarket Just Outside Marion County 34

4-3: Cumulative Costs of Walking to the Nearest Supermarket 35

4-4: Cumulative Costs of Driving to the Nearest Supermarket 36

4-5: Conversion of IndyGo Bus Routes to Raster 39

4-6: Allocation of Bus Network to Nearest Supermarket 41

4-7: Average Travel Cost to the Nearest Supermarket by Walking 43

4-8: Average Travel Cost to the Nearest Supermarket by Driving 44

4-9: Average Travel Cost to the Nearest Supermarket by Transit 45

4-10: Census Tracts Designated as Food Deserts, adapted from United States Department of Agriculture, 2013 46

4-11: Percent of Income Spent on Walking to Supermarkets Annually 48

4-12: Percent of Income Spent on Driving to Supermarkets Annually 49

Figure Page

4-13: Percent of Income Spent on Taking Transit to Supermarkets Annually 50

5-1: Marginal Effects of a Crime Intervention on the Walking Proportion of Income 64

Appendix Figures

A-1: Ordinary Least Squares Regression for Walking 77

A-2: Spatial Lag Regression for Walking.. 78

A-3: Ordinary Least Squares Regression for Driving... 79

A-4: Spatial Lag Regression for Driving 80

A-5: Ordinary Least Squares Regression for Transit.. 81

A-6: Spatial Lag Regression for Transit 82

ABSTRACT

Bailey, Andrea Leigh. M.S.C.E., Purdue University, May 2015. Low Income, Supermarket Accessibility, and the Transportation Network: A Multimodal Analysis Identifying Areas of Poor Accessibility and Intervention Strategies in Indianapolis, Indiana. Major Professor: Konstantina Gkritza.

The United States Department of Agriculture (USDA) Economic Research Service estimates that 23.5 million people live in food deserts, over half of which are considered low-income residents. Accurately defining a food desert is crucial as the designated areas can benefit from grant opportunities and funding priority. To qualify as an urban food desert, the USDA requires that at least 500 residents or one-third of the population live outside a one-mile buffer from a supermarket as well as have a median income of less than 80% of the area average or a poverty rate of greater than 20%. Approaches in the literature to identify low accessibility areas (food deserts) include simple spatial analyses, travel cost models, grocery cost models, and activity-based models. Although using cost as a measure of access is beneficial, the travel cost components are ill-defined, especially for transit. Additionally, defining food deserts as a ratio of travel cost to median household income may more accurately reflect areas with poor accessibility to healthy food by utilizing a combined measure instead of distinct income and access components.

This paper develops a cost surface for auto, transit, and walking to determine the average travel cost to the nearest supermarket for each mode in Indianapolis using Spatial Analyst in ArcGIS 10.2. Given the results from ArcGIS, spatial lag models are used to model the proportion of household income spent on traveling to supermarkets as a function of socioeconomic variables. The results show that a higher crime density, no college degree, and living outside of I-465 are all correlated with poorer accessibility to healthy food. These explanatory variables had similar effects for driving and walking, but the transit network was less sensitive to education and crime and more location-dependent. For this study, working with the police department and community to reduce crime as well as expanding the transit network are both recommended as potential interventions. Results from this analysis can provide valuable insight into the reasons behind the existence of food deserts.

CHAPTER 1. INTRODUCTION

1.1 Research Motivation

The obesity problem in the United States has been increasing in recent decades. The rate has steadily increased from 15.9% in 1995 to 27.5% in 2010. Existing research has shown a link between diet—specifically fruit and vegetable consumption—and obesity. An underlying question is whether or not residents of a community have access to supermarkets, often used as a proxy for healthy food. In some communities, convenience stores or fast food options may be more prevalent. For instance, one study found that predominantly black census block groups had a higher level of access to fast food stores (James, Arcaya, Parker, Tucker-Seeley, & Subramanian, 2014), and another study found that convenience store fresh produce provisions in African-American communities did not compensate for the lack of supermarkets in the neighborhood (Bodor, Rice, Farley, Swalm, & Rose, 2010).

In the past decade, discussion about food deserts—areas in which residents of a community lack nutritious food—has greatly increased. The United States Department of Agriculture (USDA) defines an urban food desert as a census tract that qualifies as a low-income community—meaning they have either a poverty rate of at least 20% or a median family income no greater than 80% of the area median family income—and a low-access community, defined as at least 500 people or one-third of the tract’s population living more than one mile from a large grocery store or supermarket (USDA Agricultural Marketing Service, n.d.). A supermarket is defined as a food store generating over $2 million in sales and offering a full range of food categories (Ver Ploeg et al., 2012). However, mode choice of these residents is not considered in the USDA's analysis.

Although the cause of these food deserts are unknown, they could exist for a wide variety of reasons such as supply, demand, or other market forces. Food deserts caused by supply shortages often occur in areas in which it would be undesirable for industries to locate; for instance, high-crime areas may not attract many businesses due to high shoplifting or robbery rates. Other food deserts may be caused by an insufficient demand from area residents, perhaps for financial or cultural reasons.

Because the USDA method does not account for modal constraints, this thesis proposes defining these methods by a mode-specific method. This analysis would use tools in ArcGis 10.2 by Esri to develop a cost surface for three modes: auto, transit, and walking. This measure of access from the modal analysis will then be used in a spatial and statistical analysis to examine any correlations and determine any indicators. These indicators may give some insight as to the reasons for which the food deserts exist so that proper policies or projects can be implemented to best improve access to supermarkets in that neighborhood. If successful, the increased availability of healthy foods may lower the obesity rate, thereby reducing the amount of spending on obesity-related health problems each year.

1.2 Current Issues

1.2.1 Obesity and Nutrition

Obesity is an expensive problem in the United States. It is estimated that an average of $147 billion (in 2008 U.S. dollars) is spent each year on medical-related obesity problems, and medical costs were approximately $1429 higher for those who are obese compared to those of a normal weight (Finkelstein, Trogdon, Cohen, & Dietz, 2009). A popular measure of obesity is body mass index (BMI). Although this value does not measure body fat directly, it has been shown to be correlated with other, more expensive measures of determining body fat such as dual x-ray absorptiometry (DXA) and underwater weighing (Centers for Disease Control and Prevention, 2015). BMI is calculated as

$$BMI = \frac{W}{H^2} \qquad \text{Eq. 1-1}$$

where W = *weight in kilograms* and H = *height in meters*. One drawback of the BMI method is that since it does not directly measure body fat, it fails to account for differences in body type between men and women and for people who may weigh more due to their amount of muscle; however, it is generally regarded as a good screening method, and for survey purposes the data is significantly cheaper and easier to collect. BMI results are interpreted in Table 1-1.

Table 1-1: BMI and its Weight Status Classification (CDC, 2015)

BMI	Weight Status
<18.5	Underweight
18.5-24.9	Normal
25.0-29.9	Overweight
>30.0	Obese

According to the Behavioral Risk Factor Surveillance System (BRFSS), Indiana ranked as the 8th-most obese state with 31.4% of residents classified as obese in 2012 (CDC, 2012b). Another 35.4% of the population is classified as overweight. An estimated 31.1% of residents fall into the "normal weight" category, meaning that there are approximately as many obese people in the state of Indiana as there are people of a healthy weight (CDC, 2012a). These numbers are even worse for Marion County, where the obesity rate is an astounding 33.2% (CDC, 2013). This data is self-reported and obtained by landline and cell phone surveys. Comparisons to the national average can be seen in Figure 1-1.

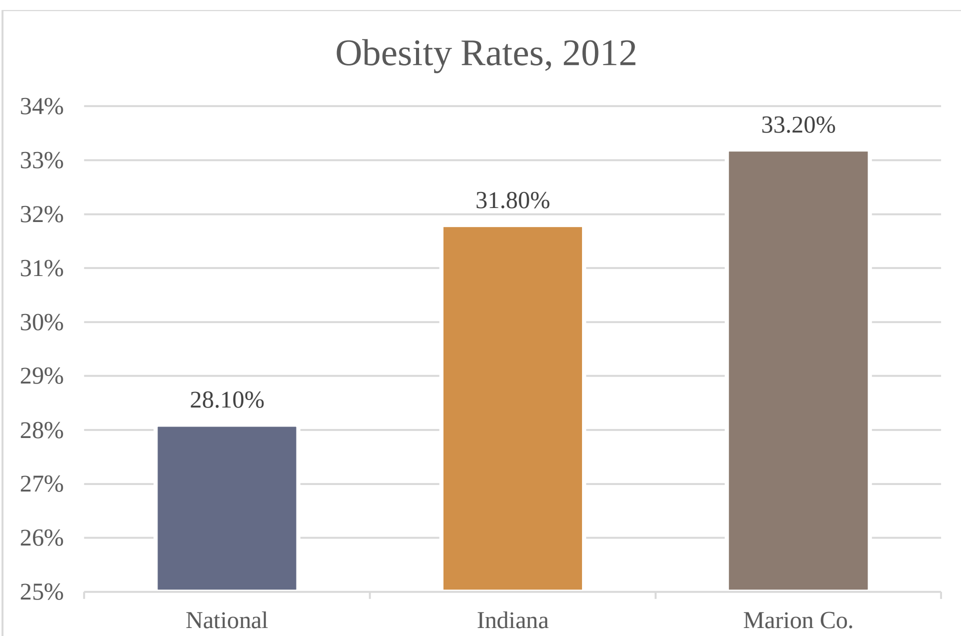

Figure 1-1: National, State, and Local Obesity Rates, adapted from CDC, 2012b and CDC, 2013

Since many studies have found an inverse correlation between fruit and vegetable intake and BMI, it is not surprising that Indiana ranks so highly for obesity but also 12th-worst in adult nutrition, defined as five or more servings of fruits and vegetables per day (CDC, 2012b). Only worsening the state's obesity problem is its lethargy; it also ranks 8th-worst for adult physical activity, defined as getting 150 minutes of moderate exercise per week. These poor health indicators are likely contributing to high healthcare spending in the state of Indiana.

1.2.2 Vehicles and Transit

Indianapolis has a reputation for being a car-friendly city. Over 91.5% of workers age 16 and over commute by automobile, but the argument could be made that a lack of transit access could partially cause that high of a number. Just 1.02% of workers in Indianapolis commute by bus. A breakdown of commuting modes in Indianapolis can be seen in Figure 1-2 (United States Census Bureau, 2013b).

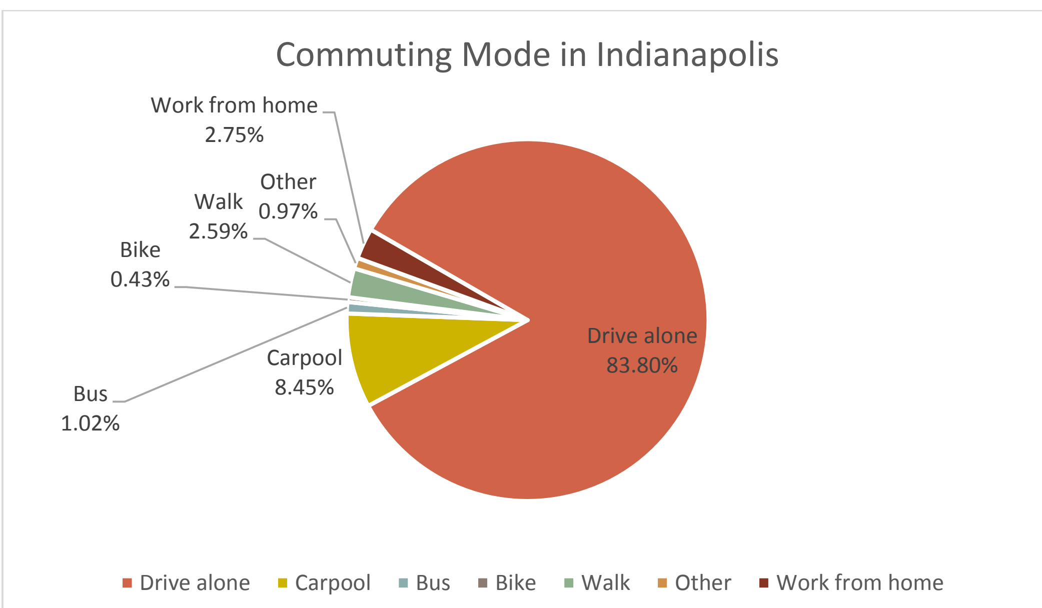

Figure 1-2: Commuting Modes of Workers in Marion County Ages 16 and Over

Additionally, a large proportion of growth in the metropolitan area has been in counties surrounding Indianapolis, and many activities outside of the immediate downtown area require a car. However, a lack of food access in urban areas is generally more of a concern for low-income groups; in a car, residents likely would not consider the time to drive to a supermarket unreasonable, but location of supermarkets relative to an individual’s home may be a significant barrier in obtaining healthy food. In fact, some studies argue that low-income households, residents in rural areas, minorities, older adults and children are most likely to have limited access to food (Coveney & O'Dwyer, 2009), (Walker, Keane, & Burke, 2010).

In general, lower income groups tend to travel shorter distances and make fewer trips, suggesting a lower level of daily mobility compared to higher-income groups (Guiliano, 2005). That said, Indianapolis ranks toward the bottom of metropolitan areas when it comes to access to transit. A Brookings report rated Indianapolis in the bottom 20 of the largest 100 metropolitan areas for access to transit, where access is defined as a ¾-mile walk from the nearest transit station to the population-weighted centroid of the block group (Tomer, Kneebone, Puentes, & Berube, 2011). This poor transit access has implications for access to supermarkets as well. Lower-income groups are more likely to take multiple bus rides and or travel long distances to access the nearest supermarket due to lack of convenient and affordable transportation, and previous studies have also shown that minority residents pay more for groceries because of the absence of major chains in their area of residence (Guiliano, 2005).

1.3 Research Objectives

This thesis has two overarching purposes. The first part of this study will identify areas with poor accessibility to supermarkets based on household income and transportation costs for different modes. After those areas are identified, a statistical analysis aims to determine correlations among travel mode, access to supermarkets, economic and demographic data, and other influences; establishing a relationship among these factors may help determine the underlying cause for these food deserts. Given the results from the statistical analysis, interventions that have the best chance of success will be recommended. The thesis then has the following two objectives:

- Determine food deserts by alternative measures to accurately reflect locations of disadvantaged residents as it relates to food access
- Determine the extent to which the transportation, social and economic environments affect the proportion of household income spent on supermarket travel costs in a census tract

1.4 Research Benefits

The end results of this study have far-reaching implications for planning or economic development organizations in the Indianapolis area. Food deserts will be better defined, especially for low-income groups that lack automobile access, which allows for better project and policy planning. The outcomes of this study could help the Indianapolis Metropolitan Planning Organization in increasing the livability of these communities lacking in access to healthy food, allowing them to complete a more mode-specific analysis.

The Indiana Economic Development Corporation could also use these results, especially if the food desert is hypothetically caused by supply-side shortages; providing incentives for supermarkets to locate in that neighborhood may attract more residents and businesses, increasing economic development. Additionally, it can be used by extension offices in order to concentrate their efforts of making healthy food more accessible in areas that need it most.

1.5 Thesis Organization

This thesis is organized into six chapters. Chapter 2, *Literature Review*, is a synthesis of previous studies' methodology for determining food deserts, and Chapter 3, *Data Description*, details the data available from various sources such as the US Census Bureau, the National Household Travel Survey, and other government agencies. Chapter 4 covers the methodology and results from the ArcGIS analysis. Methodology and outcomes from the statistical analysis along with implications are discussed in Chapter 5, *Statistical Methodology and Results*. Lastly, Chapter 6, *Conclusions*, discusses limitations and recommendations for future research.

CHAPTER 2. LITERATURE REVIEW

Although the existing literature on food deserts is expanding, there is still no clear-cut method to define those geographic areas of concern. For an urban area like Indianapolis, the United States Department of Agriculture (USDA) defines an urban food desert as a census tract that has at a median income that is 80% or less of the area family income or a poverty rate of at least 20% of the poverty level, and 500 people or one-third of the population lives at least one mile away from a supermarket or large grocery store. However, research suggests that defining food deserts may be much more complicated. This chapter presents different methods that have previously been applied to define food deserts.

2.1 Spatial Analysis

Similar to the USDA method, most research to-date used spatial methods to determine the locations of food deserts, although other considerations may vary. In a study of Toledo, OH, food deserts were determined by using the addresses of households and retailers to calculate the distance to the nearest store, excluding convenience and membership stores (Eckert & Shetty, 2011). For each block group, accessibility was estimated as an average of all of the distances to the nearest store. The process was completed for all stores and then again for only stores that were a national chain. The block group was then further examined if the following criteria were met: the nearest retailer was farther than 1 mile; the percentage of residents below poverty, the percentage of households without a vehicle, and the percentage receiving SNAP benefits were all above the city average; and the median household income was below the city average. This procedure accounts for many demographics of the block group and helps the researcher determine areas in which national chains may be unwilling to locate.

One study examined the existence of food deserts in low-income areas of New York City, particularly Brooklyn and Harlem (Gordon, et al., 2011). All retailers that sold food and beverage were included and analyzed by block but divided into categories of supermarket, bodega, and fast food. A “representative healthy food scale” was developed for bodegas; some are nutritious and culturally relevant while others contain more fast food. Census block groups were categorized by racial composition and median household income, and the availability of the variety of stores within a ¼-mile street network buffer of the population centroid was considered. A food desert index was developed to score each block group based on the types of stores to which it had access. Pearson’s product-moment correlations were used to determine bivariate relationships between demographic variables and food indicators, and the results for Brooklyn and Harlem were compared with the neighboring, affluent Upper East Side. However, the researchers note that a limitation could be that the race and income characteristics were not analyzed simultaneously.

Another study, also in New York City, considered a wide range of variables, including race, poverty, supermarket size, vehicle ownership, transit, and safety (Bader, Perciel, Yousefzadeh, & Neckerman, 2010). Supermarkets were selected for review if they had the appropriate Standard Industrial Classification (SIC) code, annual sales greater than $2 million, and at least 17 employees. Census tracts were categorized by the majority of their racial composition and the amount of their foreign-born population and then split into quartiles based on poverty rates. The street network was used to find the distance between the population-weighted center of the census tract and the nearest supermarket, but the access was weighted by the proportion of households that own a vehicle. The tract was considered to have access to transit if a stop was within 400 meters of the centroid. In order to consider safety, the investigators made a kernel density grid that contained the number of homicides and pedestrian injuries and fatalities, perhaps giving some insight into consumers' reasoning for avoiding certain areas. However, the size of the census tracts could affect the outcome of this study as well as the maximum distance chosen for walking (800 meters) and the exclusion of smaller stores.

In Edmonton, Canada, researchers examined the supermarket accessibility, particularly for high-need neighborhoods (Smoyer-Tomic, Spence, & Amrhein, 2006). A supermarket was defined as having a full range of grocery items and at least ten employees. The minimum street network distance determined the closest supermarket to the population-weighted centroid of the postal code (the smallest unit for which data were available), and the coverage method considered the number of supermarkets within 1km of the neighborhood. In addition to population data for each postal code, other information obtained at the neighborhood level included the percentage with no vehicle, percentage of people aged above 65 years, and percentage of low-income residence. The areas of low accessibility were more evident with the coverage method than the minimum distance. Many high-need neighborhoods actually had better accessibility under these measures, but six neighborhoods showed a lack of access. Zones that fell into the top quartile for those need indicators as well as the lowest quartile for spatial supermarket accessibility were determined to be food deserts.

A cross-sectional survey in South Carolina studied the relationship among the USDA-defined food deserts, healthy food retail tracts as defined by the Centers for Disease Control and Prevention (CDC), and residents' perceptions of their access to healthy food (Sohi, Bell, Liu, Battersby, & Liese, 2014). A random sample of publicly available phone numbers was used for a phone survey, asking the consumer's primary food store, store type, reasons for choosing that store, and how often they shopped. ArcGIS was used to compute distances between the chosen store and the closest store. Least squares regression was used to calculate regression coefficients, and the effect size was estimated by using the difference of the averages and the sample standard deviation. The researchers found that in general, most low-access areas traveled farther to their chosen stores and showed a higher difference in perceived access.

2.2 Activity-based Analysis

In addition to a spatial analysis, past research has used activity-based models to account for commuting patterns. A consumer's trip to work may allow more opportunities for access to healthy or low-cost foods. In Cincinnati, Ohio, data on residences and commuters were collected at the transportation analysis zone (TAZ) level in order to examine if a consumer's commute exposed them to a greater number of grocery retailers (Widener, Farber, Neutens, & Horner, 2013). A supermarket interaction potential (SMIP) score that considered commuting patterns and home-to-supermarket interaction potential (HIP) score that only considered the time budget and number of stores were calculated for the city and compared with food desert locations according to the USDA’s definition. While residents of most TAZs had more time for shopping simply basing their trip from their home locations, two zones showed improved access when considering commuting patterns by automobile. However, Widener, Farber, Neutens, and Horner (2015) found that for the same area, 43 TAZs showed improvements when considering commuting patterns by transit. When only TAZs with over 10 transit commuters were considered, 28 TAZs still showed improvements; 20 of those intersected a food desert census tract.

Additional research in Tallahassee, Florida used time geographic density estimation (TGDE) in order to examine the effects of consumers’ daily trips on their food environments (Horner & Wood, 2014). Actual travel data was unavailable, so synthetic travel records of eleven adults from a previous study were used. Supermarkets and warehouses were weighted four times as much as specialty and other small stores based on the difference in sales, and convenience stores were eliminated. The authors then used TGDE to approximate the likelihood of a vehicle being at a particular location given the consumer’s travel time budget and intermittent GPS locations. From that information, food accessibility measures for various time budgets were calculated based on the likelihood that the consumer could deviate from their shortest path. The accessibility scores were similar across varying levels of travel time budgets, suggesting that the accessibility scores are more dependent on the geographic distribution of food locations than the consumer’s activity patterns.

2.3 Travel Cost Analysis

A common method for accounting for varying levels of income as well as different travel modes in mode choice studies is the travel cost method. Hallett & McDermott (2011) sent a survey asking residents of Lawrence, Kansas, the distance to their favorite grocery store and the mode they used to travel there. The authors used a cost surface to determine the lowest-cost path between each raster cell and the closest full-service grocery, defined as a retailer with more than 30,000 square feet. Spending ten percent of the average grocery budget for a Midwest consumer on travel costs was assumed to be the threshold for an underserved area. Although they found that no resident with a car would have limited access, residents without a car would be underserved in certain areas of the city.

Researchers in Melbourne, Victoria, Australia also used a cost surface to estimate the relationship between access and socioeconomic index (Burns & Inglis, 2007). The three major Australian supermarket chains were used as a proxy for healthy food, and fast food chains with at least ten franchises in Australia were used as a proxy for unhealthy food. A raster was created in ArcGIS, with the travel time being the cost to travel each cell in a vehicle, by bus, or on foot. Using the population density of their census collection districts, the percentage of people within eight minutes of either a fast food restaurant or supermarket was calculated for each mode. Additionally, areas where a supermarket was closer were differentiated from areas where a fast food outlet was closer. Residents of a higher socioeconomic index had better access to supermarkets, while those of a lower socioeconomic index had better access to fast food. This relationship between access and socioeconomic index was found to be statistically significant using analysis of variance techniques (ANOVA).

2.4 Other Analysis Methods

Some studies combined different methods of determining food deserts. In Portland, Oregon, researchers hypothesized that although grocery stores may seem plentiful in an area, their cost could make them inaccessible to those of lower income (Breyer & Voss-Andreae, 2013). They designed a "healthy foods market basket survey" based on the Thrifty Food Plan, the basis for the Supplemental Nutrition Assistance Program (SNAP) through the USDA, more commonly known as the food stamp program. Retailers under consideration included supermarkets and other grocery stores, defined as a location that sells at least ten different fresh produce items. Their findings at various stores in the city were combined with the SNAP income and budget levels to create an "affordability index." The street network distance was measured from the block group centroid to the nearest store and to the nearest low-cost store (defined as having an affordability index less than 1) and aggregated to give an average score for that census tract. The difference between those two scores indicated the severity of the "food mirage," meaning that stores were located within a reasonable distance but a cost barrier existed for low-income groups. Some census tracts were eliminated from consideration since they had a low-cost store within one mile, and tracts in the top quartile of income were assumed to be able to drive to any location and therefore eliminated as well. A spatial lag model using the average score over the nearest five neighbors was developed to attempt to capture a relationship between demographics and food mirage locations. Three geographic areas of Portland were modeled, and two different regression analyses were conducted for each region: one using a dependent variable of potential food mirage distance and the other with a dependent variable of distance to the nearest grocery store. An advantage of this study is that it addressed some demand components for the existence of food deserts, but walking is the only mode considered and preferences (brand, cultural, etc.) are not taken into account.

CHAPTER 3. DATA DESCRIPTION

This chapter will detail the sources of the data available for this geographic and economic analysis. A rationale for choosing Marion County as the location for this investigation is included, as well as a description of the U.S. Census Bureau's geographic, economic, and demographic information, which is the primary source of data. Additionally, other sources are discussed, including IndyGo's General Transit Specification Feed reference, the 2009 National Household Travel Survey, the North American Industry Classification System, and the Uniform Crime Report.

3.1 Geographic Selection

Marion County was selected for several reasons. First, it includes the City of Indianapolis as well as several other small municipalities. Marion County and the City of Indianapolis have a unigovernment, meaning the county and city governments are combined, so the county level was chosen for ease of analysis. Due to its urban setting, a large amount of data is available, and many census tracts are small enough that they can be used as the geographic basis for analysis. Marion County contains 224 census tracts. Although census block groups are smaller and may provide a more accurate spatial analysis, many economic characteristics such as median income, personal vehicles available, poverty level, and food stamp participation are not available at a more disaggregate level. Additionally, as discussed in the introduction, Indianapolis has relatively poor transit systems, so low-income groups have the potential to be more at a disadvantage when compared to a transit-friendly city such as San Francisco, Washington, or Chicago.

3.2 Data Sources

3.2.1 TIGER/Line Files

The primary source of data for this project is TIGER/Line files from the United States Census Bureau. TIGER stands for Topically Integrated Geographic Encoding and Referencing and alludes to the fact that the files combine geographic information with other quantitative data. For this purpose, census tracts were chosen as the level of analysis because of their size in urban areas and the amount of information available at that scale. Several versions of these TIGER/Line files are available, but for this project, TIGER/Line files with Selected Demographic and Economic Data were used. Although extensive tables with this information can be downloaded from the census website, a useful attribute of these integrated files is that the census data is combined with shapefiles that can be used with GIS programs, specifically ArcGIS by Esri. The shapefiles contain the geographic components of the census tracts, such as legal boundaries, on a geographic coordinate system. Other files that can be obtained through the database include a road network of primary and secondary streets, which was used to complete a network analysis. Information was also obtained for each census tract in Marion County from the 2008-2012 American Community Survey 5-Year Estimates. These variables include demographic data such as race and age; economic information such as median income and percentage of residents beneath the poverty level or participating in SNAP; and vehicle and commuter information such as mode choice when commuting to work and the number of personal vehicles available to workers. Descriptive statistics of some of these tract characteristics are displayed in Table 3-1 below.

Table 3-1: Descriptive Statistics of Selected Socioeconomic Variables

Variable	***Mean or %***	***Std. Dev.***	***Obs.***	***Maximum***	***Minimum***
Median household income	$43,631	$19,929	224	$113,576	$14,299
Households below the poverty level	612	519	222	4,400	61
Households that accept food stamps	175	134	222	677	0
White residents	87.3%	--	223	100.0%	20.3%
Black residents	4.6%	--	223	68.8%	0.0%
Hispanic residents	3.2%	--	223	26.7%	0.0%

Other shapefiles necessary for this analysis were the roads. The roads are coded by county, interstate, common name, state-recognized, U.S., and other.

3.2.2 IndyGo: Indianapolis' Bus System

Another important component of this analysis is the existing transit network in Indianapolis. A set of text files were obtained from IndyGo's developer website called the General Transit Feed Specification (GTFS) Reference (IndyGo, 2014). Many transit networks use this reference system, which contains information about the location of stops, timing of stops, route paths, and calendars, and it is frequently used by mobile application developers. The files available from IndyGo can be found in Table 3-2.

Table 3-2: GTFS Reference Files Available for IndyGo, adapted from Google Developers, 2015

File name	***Definition***
agency.txt	One or more transit agencies that provide the data in this feed
calendar.txt	Dates for service IDs using a weekly schedule. Specify when Service starts and ends, as well as days of the week when service is available
calendar_dates.txt	Exceptions for the service IDs defined in the calendar.txt file
routes.txt	Transit routes, or groups of trips displayed to riders as a single service
shapes.txt	Rules for drawing lines on a map to represent a transit organization's routes
stop_times.txt	Times that a vehicle arrives at and departs from individual stops for each trip
stops.txt	Individual locations where vehicles pick up or drop off passengers
trips.txt	Trips for each route. A trip is a sequence of two or more stops that occurs at a specific time

An employee at Esri created additional tools and made them available for download in order to aid in this transit network analysis (Morang, 2014). These tools can be added to the ArcGIS toolbox and use the files available from the IndyGo GTFS. One tool called *Display GTFS Route Shapes* uses the trips, routes, and shapes text files to map the transit network. Other tools include *Better Bus Buffers*, which counts trips for individual routes, in polygon buffers around stops, and at points and stops, as well as *Add GTFS to Network Dataset*, which helps prepare the transit system for analysis using *Network Analyst*.

3.2.3 National Household Travel Survey

Administered periodically by the Federal Highway Administration (FHWA), the National Household Travel Survey (NHTS) can provide insights into drivers' behavior and was formerly called the Nationwide Personal Transportation Survey. Survey data is collected through a computer-assisted telephone interview, and participants are chosen by a list-assisted random digit dialing process given that they are not living in a prison, in a medical institution, or in barracks on a military base (Federal Highway Administration, 2011). Data collected includes trip purpose, mode, duration, and time of day as well as vehicle characteristics. The most recent survey is from 2009; information such as average number of vehicles per household, person-miles, person-trips, vehicle-miles, and vehicle trips is then extrapolated to census tracts in that metropolitan area through a process called transferability with the help of the Bureau of Transportation Statistics (BTS), using information given by 5-year ACS estimates from 2007-2011 (two years before and after the survey year). Descriptive statistics of the transferability results are listed in Table 3-3 below.

Table 3-3: NHTS Data Descriptive Statistics

Variable	***Mean***	***Std. Dev.***	***Obs.***	***Maximum***	***Minimum***
Person miles traveled	55.74	10.98	222	31.63	91.84
Person trips	8.58	1.24	222	5.24	11.91
Vehicle miles traveled	37.97	9.24	222	18.46	62.85
Vehicle trips	5.22	0.93	222	2.91	7.51
Household vehicles	1.58	0.29	223	0.77	2.3

3.2.4 Supermarket Locations

Information about the locations at which people can obtain food is crucial to this analysis. Stores were selected by their code in the North American Industrial Classification System (NAICS), which government agencies use to classify all businesses for the purpose of analyzing and publishing data related to that market sector. Stores have one primary code and can have several secondary codes. Using the code 445110 for “supermarket and other grocery (excluding convenience) stores,” a reverse lookup was done on a website called ReferenceUSA to find businesses with that particular primary NAICS code. Convenience stores were excluded for two primary reasons. First, since the USDA does not include convenience stores when determining food deserts, excluding them is more consistent with previously defined methods. Additionally, the majority of convenience stores do not sell a wide range of fresh produce, and given the relationship between nutrition and obesity, convenience stores may overstate the availability of healthy food. Further, although some convenience stores may sell fresh produce, it would be costly to visit each store individually to determine the type of food; keeping the cost of obtaining data low makes this method more repeatable for other regions. A total of 133 stores were found. Along with the address of the business, other available data included latitude and longitude, estimated sales per year, and an employee range. Stores were found in Marion County as well as the surrounding counties to account for stores near the county lines that may still be within a reasonable distance of Marion County residents.

3.2.5 Safety

Lastly, some businesses may choose their ideal location based on the safety of the environment; in other words, high-crime areas will occasionally deter investors because the risk is seen as too high. The safety of an area may also be a factor in consumers' willingness to use certain forms of transportation, such as walking, in order to obtain goods. In order to investigate these possible correlations, crime data was collected from the Uniform Crime Report (UCR) published by the Indianapolis Metropolitan Police Department (IMPD). The UCR is organized by the Federal Bureau of Investigation and provides a consistent manner in which local jurisdictions can report violent and property crimes. This report includes information such as the type of crime, the northing and easting in feet based on Indiana State Plane East coordinates, and time of day (Indianapolis Metropolitan Police Department, 2012). Due to the unigovernment system in Indianapolis, though, the IMPD districts do not include a few "excluded cities" that instead have their own police department (see Figure 3-1). These excluded cities include the City of Lawrence, the City of Beech Grove, and the Town of Speedway. Although information was unavailable for the Lawrence Police Department, the Indiana University-Purdue University Indianapolis Police Department, and the Indianapolis Airport Authority, crime counts for each type were published by the UCR for the Speedway and Beech Grove Police Departments (Federal Bureau of Investigation, 2012).

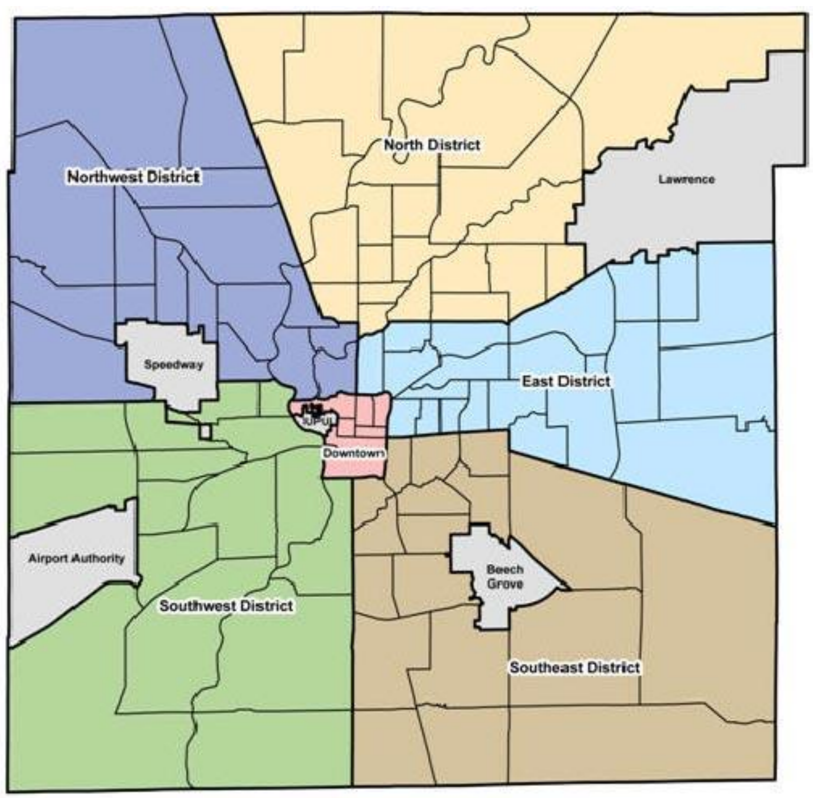

Figure 3-1: Indianapolis Metropolitan Police Department Districts (IMPD 2010)

3.3 Preliminary Data Analysis

The NHTS and the TIGER/Line files with selected demographic and economic data already contained information specific to census tracts; however, much of the other acquired data still required preliminary processing. These steps were required to obtain information about the community and built environment so that possible correlations could be found and proper recommendations made. The first step was to project the data from its original geographic coordinate system (GCS North American 1983) to a projected coordinate system (NAD 1983 Indiana State Plane East FIPS 1301, meters) to complete any analysis related to area, such as adding buffers around supermarkets and bus stops or calculating kernel densities.

3.3.1 Transit Network

The GTFS Reference provided from IndyGo contains the latitude and longitude of each bus stop in the network, so first, the "Display XY data" tool was used to create a shapefile and plot the locations across Marion County. This shapefile was spatially joined to the census tract shapefile, giving the number of bus stops per each census tract. A field was also created to calculate the density of bus stops in each census tract. Lastly, the *Display GTFS Route Shapes* tool used the stop locations and the route order to create unique polylines for each route.

3.3.2 Supermarkets

First, the number of stores returned from the reverse NAICS code lookup required some filtering. These businesses were refined by the amount of sales and the number of employees. Stores with fewer than five employees or less than $1 million in sales were eliminated from the analysis. These criteria are similar to the U.S. Department of Agriculture's definition of supermarket, which uses a threshold of $2 million in sales; it is also assumed that stores of such a small scale do not have a selection of healthy food that is comparable with other stores. However, it is important to note that ethnic stores were not eliminated unless they did not meet the previously mentioned criteria. The remaining stores were then cross-checked using a search engine and Google Maps. If the store was determined to not be a full-service grocery (for example, a spice store) or not found to exist in that location, then it was eliminated. In total, 41 stores were removed from consideration. Then, similar to the preliminary analysis of the bus stops, the supermarkets that meet the criteria undergo the spatial join process to calculate the number of stores per census tract.

3.3.3 Crime Locations

The preliminary analysis of the crime locations is also similar to that of the bus stops. The location data from the IMPD Unified Crime Report is given in Indiana State Plane East northing and easting in feet, so it had to be transformed using the same coordinate system but in meters. Because only the total number of crimes, not their locations, were available in the Speedway and Beech Grove jurisdictions, a new polygon layer was added in which the *Random Points* tool in ArcGIS was used to represent the number of crimes. After the locations were converted and plotted, the number of crimes per census tract was calculated using the spatial join function; a crime density was calculated based on those results and using the area of the tract. The crime density per census tract is shown in Figure 3-2.

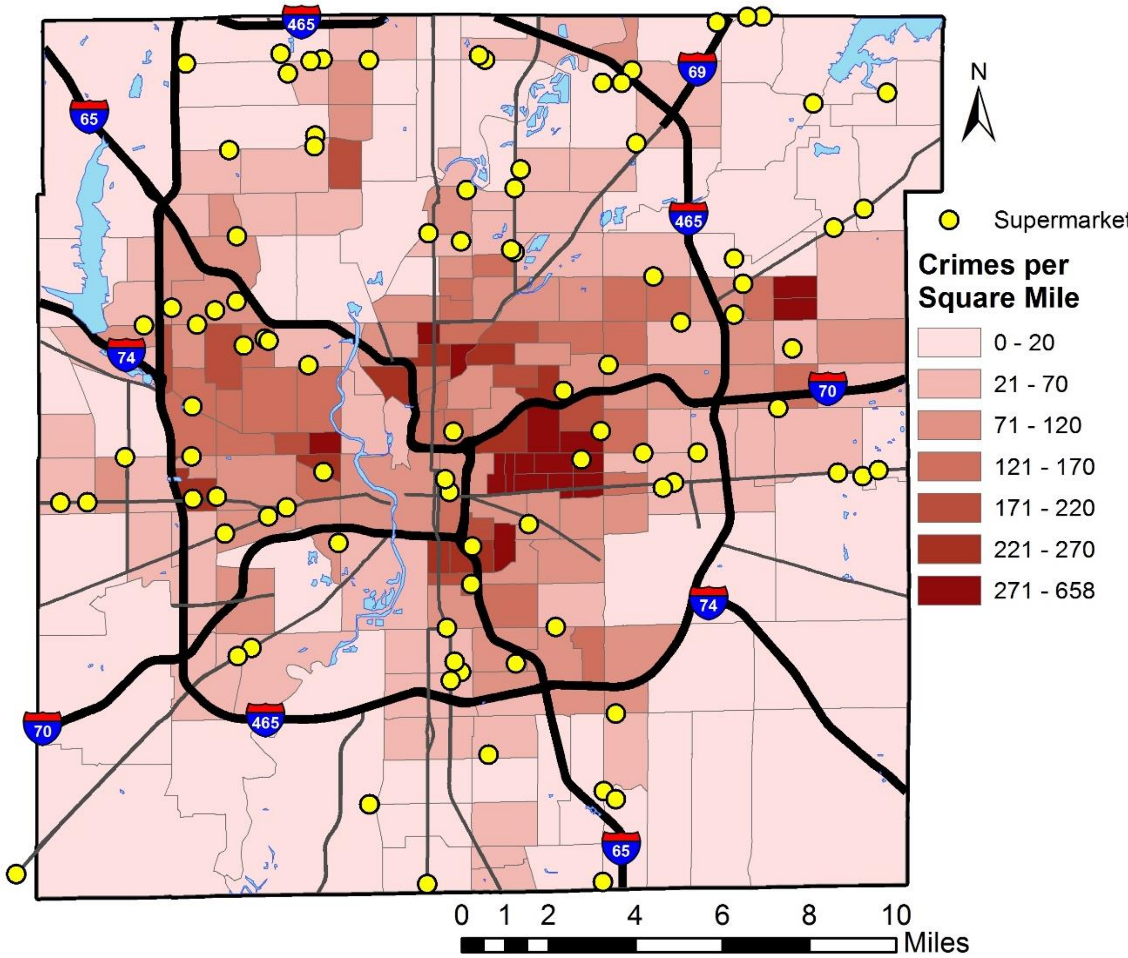

Figure 3-2: Crime Density in Crimes Per Square Mile for Each Census Tract

A raster of crime density was also developed using the kernel density tool, which uses a quadratic function to value events close to the cell more highly and gradually decreases to zero at the radial boundary. A 30 meter-by-30 meter cell size was used as well as a 1609-meter search radius and can be seen in Figure 3-3. This method more accurately represents the crime density in the area because it is not affected by the census tract boundaries.

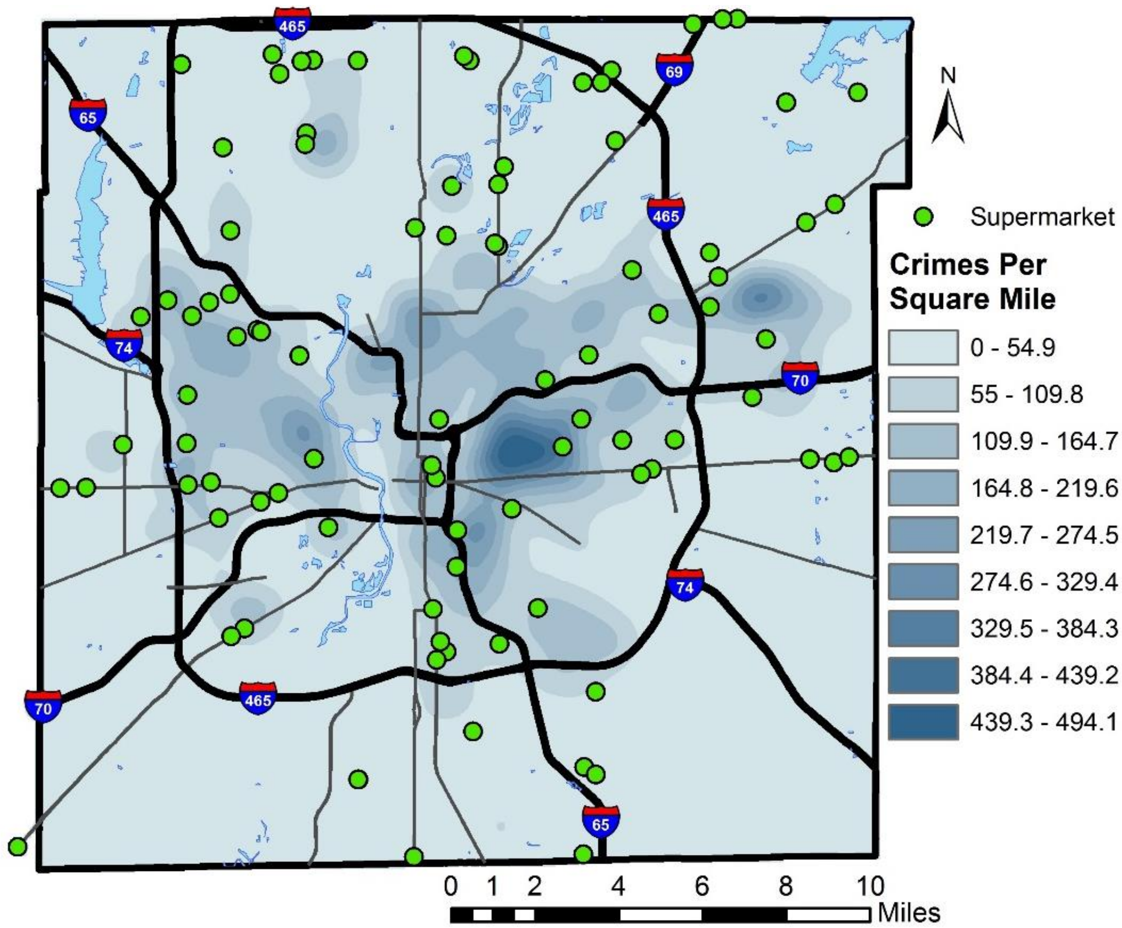

Figure 3-3: Crimes Per Square Mile for Each 30m x 30m Cell using a 1-Mile Search Radius

CHAPTER 4. ARCGIS METHODOLOGY AND RESULTS

This chapter details the methodology utilized for the spatial analysis. The first half of this chapter outlines the methods and tools used in ArcGIS to determine residents' access to food based on transportation cost by walking and driving, and the second half of this chapter expands on the walking and driving methodology to complete a multimodal transit analysis.

4.1 Background Information

Although several methods exist to determine food deserts, this study builds on existing methodology and expands on some limitations of other studies. For example, the study in Lawrence, Kansas does not include a value of time for driving, and the study in Melbourne, Australia does not include the cost of operating a motor vehicle and considers transit frequency as the primary measure of access (Hallett & McDermott, 2011; Burns & Inglis, 2007). The overarching purpose of this study is to identify areas that are disadvantaged as it pertains to access to healthy food as well as best-practice interventions based on possible correlations. Costs for multiple travel modes are considered, which are used in the next section and modeled against other available data to find statistically significant correlation factors. By examining correlations, interventions can be strategically selected and have a higher potential for success.

Another important consideration is the time period over which the area is studied. At the inception of this analysis, the most recent year for which data was available was 2012, so all data published yearly references that year. The only exception is the 2009 NHTS, since those surveys are only completed periodically. The cost calculations in ArcGIS described in the following section could easily be repeated in the future using updated files available at that time.

4.2 Walking and Driving Analysis

ArcGIS has many powerful tools within the Spatial Analyst toolbox. As mentioned previously, shapefiles referencing a geographic coordinate system or created using latitude and longitude will need to be converted using the *Project* tool.

4.2.1 Converting the Road Network to a Raster

In order to use the cost distance tool in ArcGIS, the road network must be converted to a raster, which is accomplished using the *Polyline to Raster* tool. The road network shapefile contains a field labeled MTFCC, an abbreviation for MAF/TIGER Feature Class Code, which is used for this step. The Census Bureau uses these codes to classify features or objects in GIS shapefiles. Because the *Polyline to Raster* tool classifies the cell based on the primary line type that passes through it, separate walking and driving road files will need to be used to ensure a complete network. For instance, since walking is prohibited on interstates, a cell that included both local roads and interstates could be classified as an interstate road type and therefore seen as a gap in the walking network. Although the database of feature class codes is large, the ones used in this shapefile are listed in Table 4-1.

Table 4-1: Road Type Classifications (United States Census Bureau, 2015)

MTFCC	***Feature Class***	***Description***
S1100	Primary Road	Generally divided highways distinguished by interchanges
S1200	Secondary Road	Main arterial
S1400	Local Neighborhood Road, Rural Road, City Street	Paved non-arterial road, usually 2-lane
S1500	Vehicular trail	Unpaved dirt trail
S1630	Ramp	Entry to or exit from limited access road
S1640	Service Drive	Gives access to structures along a limited-access highway
S1710	Walkway/Pedestrian Trail	Restricted from vehicular traffic
S1730	Alley	Service road generally at the rear of buildings
S1740	Private service vehicle road	Privately maintained for service purposes
S1750	Internal U.S. Census Bureau use	Internal U.S. Census Bureau use
S1780	Parking Lot Road	Main vehicular route through a paved parking area
S1820	Bike Path or Trail	Restricted from vehicular traffic

4.2.2 Reclassifying the Road Raster as a Cost Raster

The road raster is used as a basis for converting each cell to a cost value. The primary costs of travel are a value of time and operating costs; both walking and driving costs will account for the value of time, but only driving has an operational cost associated with it. The cost of traversing a unit distance is given by the following formulas:

$$TC_d = \frac{\frac{FMW}{s} \times 100 + c}{1609.34} \qquad \text{Eq. 4-1}$$

$$TC_w = \frac{FMW \times 100}{w \times 1609.34} \qquad \text{Eq. 4-2}$$

where TC = transportation cost by mode in cents per meter; FMW = federal minimum wage, currently at $7.25; s = driving speed in mph; c = cost of operating a motor vehicle in cents; w = walking speed, assumed to be 3.0 mph; and 1609.34 is a conversion factor between meters and miles. It is not necessary to multiply by the length of the cell because ArcGIS accounts for that distance and allows the user to easily change the cell size. According to the Bureau of Transportation Statistics, the average cost per mile of operating a motor vehicle was 60.8 cents per mile in 2012 (Bureau of Transportation Statistics, 2014) This cost is based on driving 15,000 miles per year and includes variable costs such as gas, maintenance, and tires as well as fixed costs such as license, registration, insurance, and depreciation.

The *Reclassify* tool was then used to convert the cell's road type to its cost. Although the above equations give transportation costs in cents, it was necessary to convert them to thousandths of cents because the tool only accepts integers as input. Because of the differing equations chosen to represent the transportation cost for each cell, each mode required the creation of separate cost rasters. The assumed travel speeds by automobile for each road type can be found in Table 4-2 as well as the cost associated with traversing those cells. Additionally, the road types representing pedestrian paths and bike paths were excluded from the driving analysis by reclassifying them as NoData. The driving network raster is displayed in Figure 4-1.

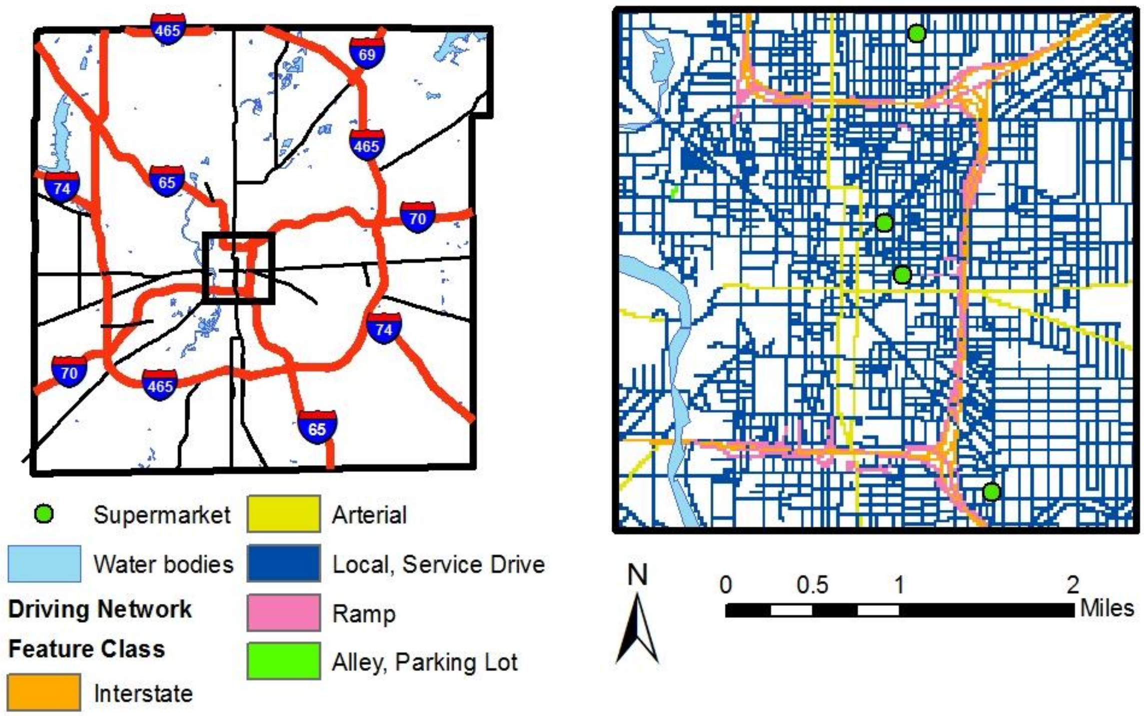

Figure 4-1: Inset of the Driving Raster by Road Type

Table 4-2: Driving Speeds and Costs by Road Type

Feature Class	Speed Limit (mph)	Drive Cost (0.001 cents/m)
Primary Road	55	45.66
Secondary Road	40	48.73
Local Neighborhood Road, Rural Road, City Street	30	52.49
Vehicular trail	--	NoData
Ramp	25	55.49
Service Drive	30	52.49
Walkway/Pedestrian Trail	--	NoData
Alley	10	82.52
Private service vehicle road	--	NoData
Internal U.S. Census Bureau use	--	NoData
Parking Lot Road	10	82.52
Bike Path or Trail	--	NoData

For walking, all travel speeds were assumed to be the same; however, the pedestrian and bicycle paths were included and interstates were excluded from the analysis. All other feature class types that were excluded for driving were also excluded for walking. Using the above equations, the value for walking was approximately 0.15 cents per meter, which is roughly two to three times the values for driving.

4.2.3 Finding the Cost to the Nearest Supermarket

After the cost to cross each cell has been determined, cumulative costs to the nearest grocery store can be calculated. Methods using Euclidean distance may not account for bodies of water or areas in which there are no roads; a benefit of using a travel cost method is that it does not ignore such barriers. Additionally, since every residence must somehow be connected to a road network, large fields or industrial areas with few roads do not have a large effect on the analysis while neighborhoods with a higher road density are more heavily weighted.

After the supermarket locations were finalized as outline in Chapter 3.3.2, the final number of stores was 92. However, one of the supermarkets included is just outside the southwest corner of Marion County, and a polyline was extended to allow households access in that census tract, which can be seen in Figure 4-2. This southwestern-most tract has a concentration of residential land but is surrounded by industrial land and the Indianapolis International Airport, so excluding that supermarket from analysis caused the *Cost Distance* tool to not accurately reflect true costs in that area.

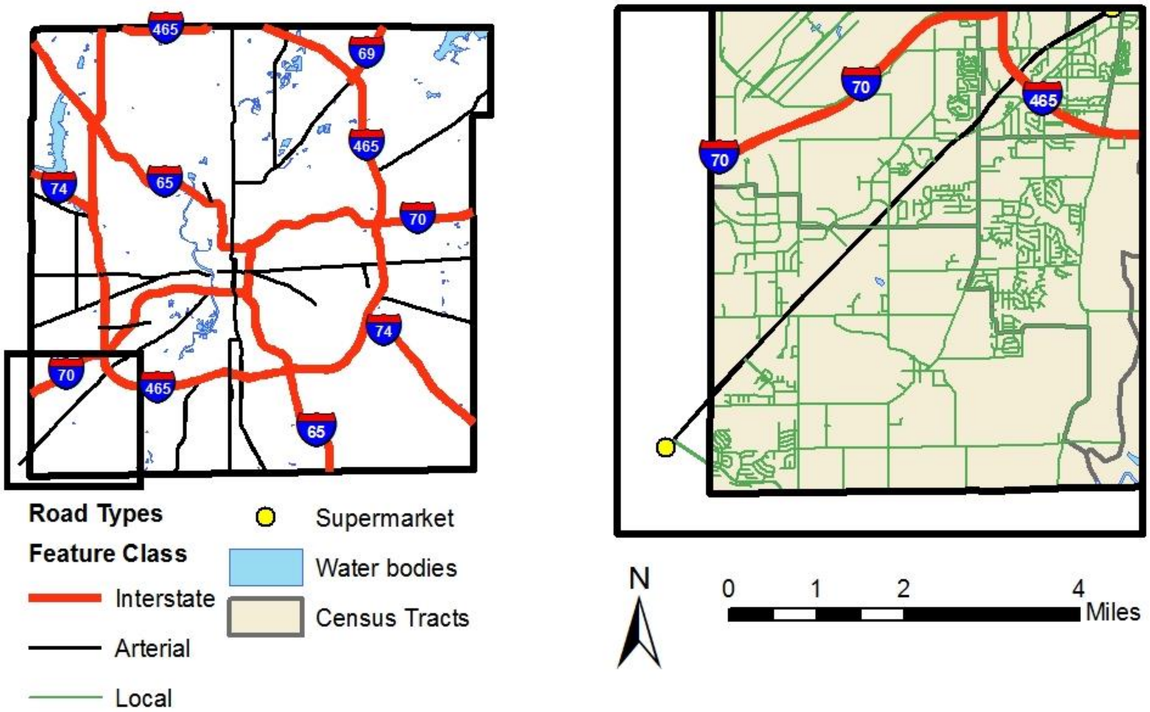

Figure 4-2: Accounting for a Supermarket Just Outside Marion County

Once the supermarket selections had been finalized, it was necessary to ensure that they were included in the network created by the road raster. While some stores had listed geographic coordinates close enough to the road centerline that they were contained in the same cell as part of the road network, other businesses were far enough off the road that the analysis tools did not recognize them as being accessible by this network. In order to fix this issue, the supermarkets were moved to be immediately adjacent to the nearest street.

The *Cost Distance* tool was then used to determine the least accumulative costs to the nearest supermarket for each cell in the network. When running the tool, *input raster or feature source data* is the supermarket shapefile, and the *input cost raster* is the cost raster created in the previous step. It allocates each cell to the supermarket to which the transportation cost is lowest. This procedure is repeated for each mode, respectively; the costs for walking and driving can be seen in Figure 4-3 and Figure 4-4.

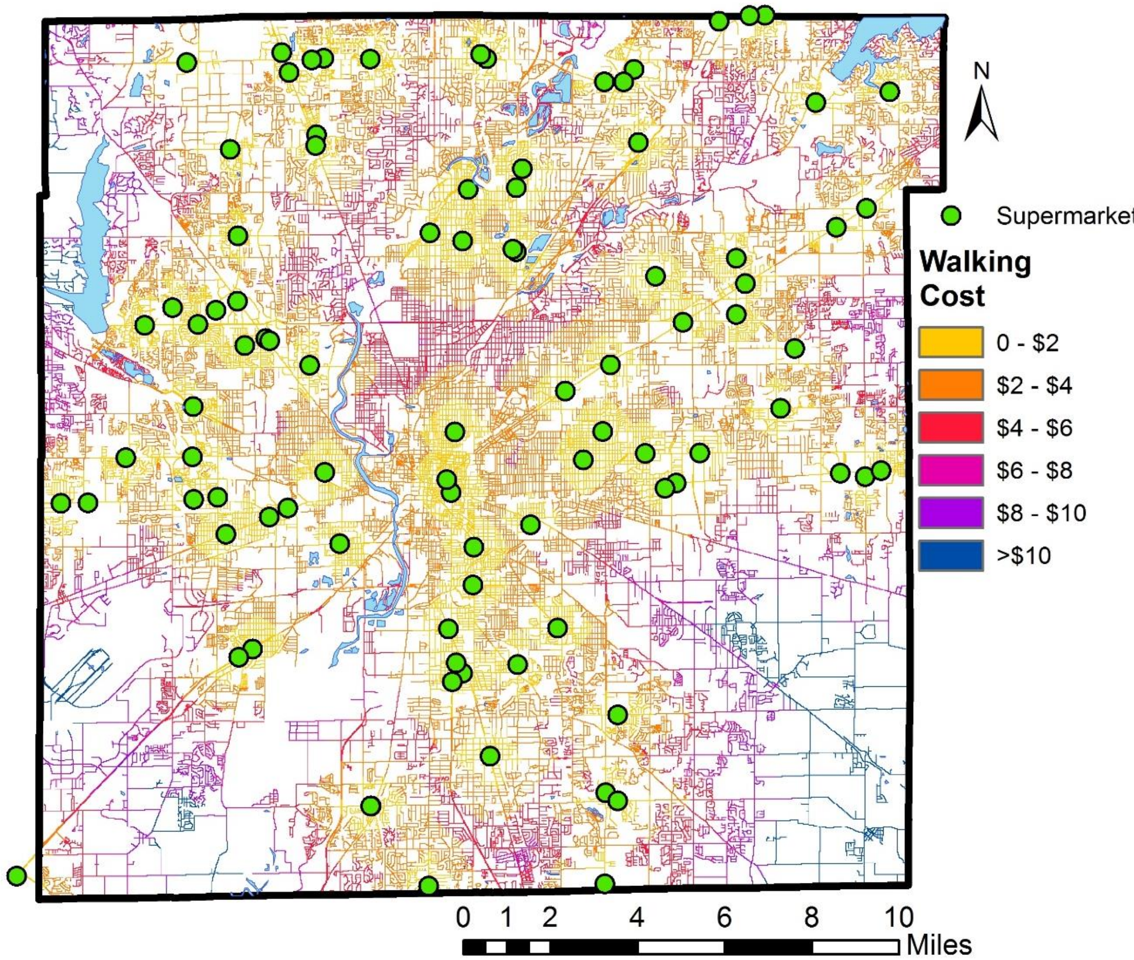

Figure 4-3: Cumulative Costs of Walking to the Nearest Supermarket

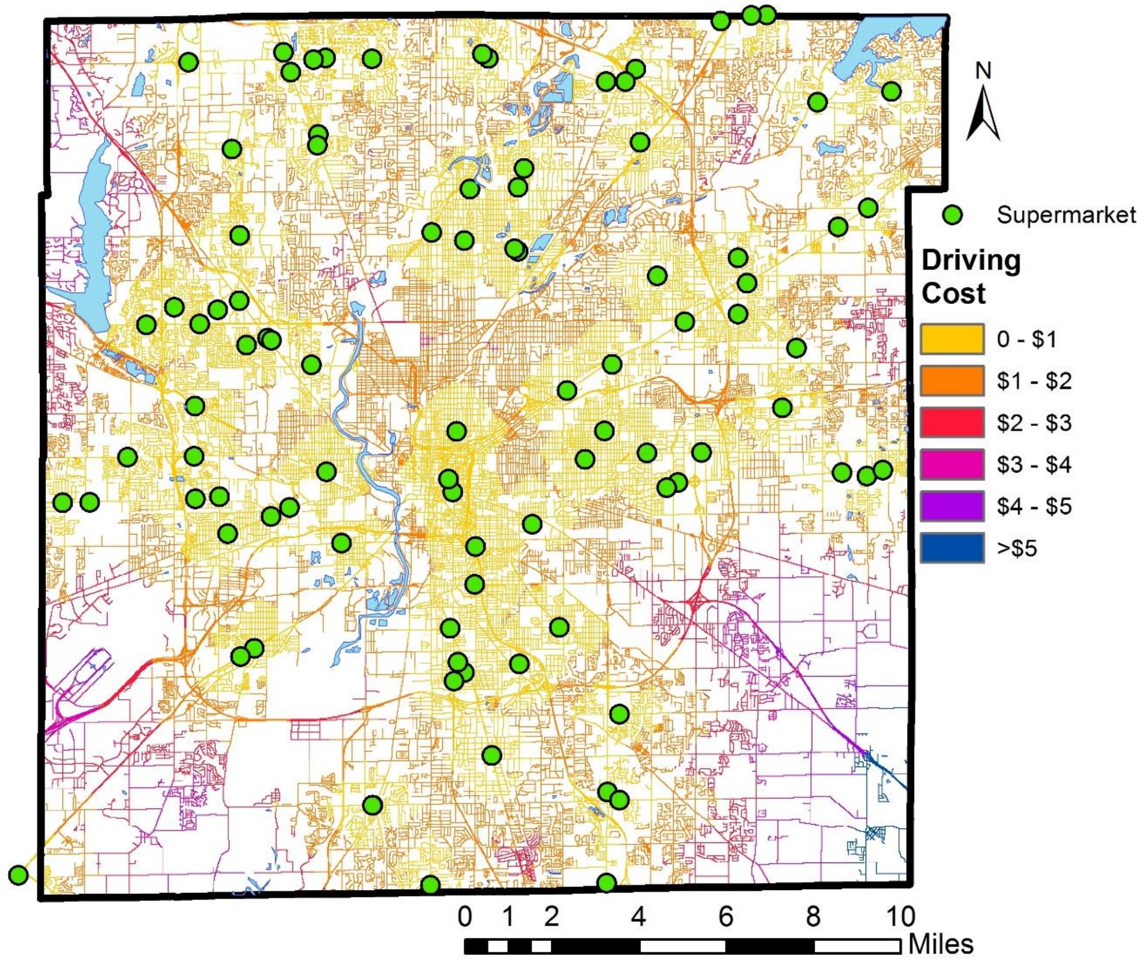

Figure 4-4: Cumulative Costs of Driving to the Nearest Supermarket

4.2.4 Determining Statistics for Each Census Tract

Once the travel cost to the nearest supermarket is completed for each cell, the data is aggregated using the *Zonal Statistics* tool. The census tracts shapefile is used for the feature zone data, the unique census tract identifier is used for the *zone field*, and output from the cost distance tool is used as the input value raster. Mean is chosen as the statistics type. However, although this does generate unique average values for each census tract, the results are displayed only visually and do not provide those values in output form for further analysis form.

Since the *Zonal Statistics* tool itself does not have a table of numerical output, further mathematical analysis can be completed using the *Zonal Statistics as Table* tool. Using the same steps above, raster statistics of mean, standard deviation, minimum, and maximum can be aggregated by census tract boundaries. Although there is no map output, the table can be joined to the attribute table of the census tracts for visual purposes or exported to a file for further analysis. Figures from this step showing the average travel costs can be seen in the results section of this chapter.

4.3 Transit Analysis

Because of the multimodal nature of transit, this analysis was significantly more complicated than the walking or driving. Walking or driving deals with specific origin-destination pairs, while transit requires three: walking to a bus stop, taking the bus to the stop nearest a supermarket, and walking from the bus stop to the supermarket. Because of the three separate segments of transit travel, this analysis is completed in three distinct steps in reverse order. However, similar to the driving and walking analyses, one limitation to note is that the Spatial Analyst tools do not account for route transfers or the specific bus route stop sequence, only the general shape that the bus network follows.

4.3.1 Preliminary Steps

Before the major portion of the analysis begins, certain rasters need to be created to make the following steps easier. The primary issue at hand is that since the bus stop coordinates are located off of the road, they are not necessarily contained in the same raster cell as the walking or transit networks due to the grid pattern. These cells with bus stops need to be added to the walking and bus raster networks in order to ensure that no bus stops are ignored in the analysis.

The *Polyline to Raster* tool is used to convert the transit lines into 30-meter by 30-meter cells, and the *Point to Raster* tool is used to determine each cell that contains at least one bus stop. An important note when adding two rasters is that a cell with a value plus a cell with NoData will equal a cell with NoData. Each of these rasters must be reclassified so that each NoData cell is equal to zero in order for them to be combined; otherwise, the resulting raster would only contain values for any cell that included values for both the bus stops and the bus lines. The transit line and bus stop rasters are then added together and reclassified by the cost per unit distance of traversing each cell, which is given by:

$$TC_b = \frac{FMW \times 100}{b \times 1609.34} \qquad \text{Eq. 4-3}$$

where TC_b = travel cost by bus in cents per meter, *FMW* is the federal minimum wage in 2012 dollars, and b = the bus speed, assumed to be 30 mph. During that reclassification process, all cells with a sum of zero must be reclassified as NoData in order to be ignored in the cost distance analysis. Additionally, in order to input this as an integer form as the *Reclassify* tool requires, the cost per unit distance will need to be converted to thousandths of cents.

The same process is repeated to add the bus stop cells to the walking network so that pedestrians can access them; that raster is reclassified by the cost of walking given in the previous section. Note that bus fare is not included in this equation because IndyGo uses a fixed rate; it will be added at the end of the analysis. The total bus network is displayed in Figure 4-5.

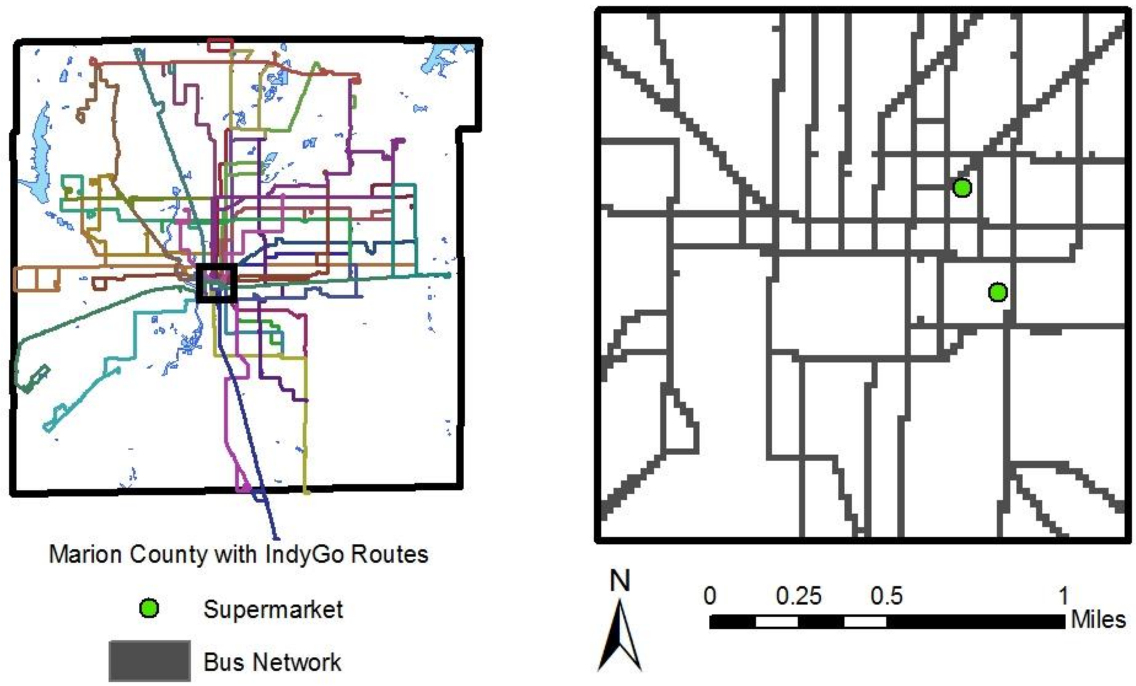

Figure 4-5: Conversion of IndyGo Bus Routes to Raster

4.3.2 Step 1: Distance from the Bus Stop to the Nearest Supermarket

The last leg of the trip is calculated first so that cost can be allocated to other cells in the network. First, the *Near* function is used to find the nearest bus stop to each supermarket. This results in 87 bus stops for 92 supermarkets, as two bus stops serve two stores and one bus stop serves three. A new layer is created with these selected bus stops for further analysis and will further be referred to as “supermarket bus stops.” The *Point to Raster* tool is again used to convert this new layer of supermarket bus stops using the “count” field in the file. The resulting raster will be comprised of 87 cells with a value of one and the remaining cells with NoData.

Because consumers will have to walk from the supermarket bus stop, the walking network with the added bus stops created in the previous section is used as input for the

Cost Distance tool, with the feature source being the supermarkets. Then, by adding this raster to the one with the supermarket bus stops and subtracting one (the supermarket bus stop raster's original value), the cost of walking from each of those stops to the nearest grocery can be determined. All other cells in the cost distance raster will cancel out since the other raster's corresponding cells are classified as NoData.

4.3.3 Step 2: Distance from Each Bus Stop to the Nearest Supermarket Bus Stop

The next step is to calculate the cost of travel time while on an IndyGo bus. Each cell in the transit network needs to be allocated to the nearest supermarket bus stop on a basis of cost, which can be accomplished using the *Cost Allocation* tool. The raster with the cost of walking from the supermarket bus stops to the stores is used as the input raster, but it first must be rounded to the nearest integer per the requirements of ArcGIS with the *Int* tool. It is not expected that this will have a significant effect on the analysis since values are already in thousandths of cents. Once this is complete, the new integer raster values are distributed through the cost allocation process to cells throughout the transit network. The results of the cost allocation can be seen in Figure 4-6, where each color in the map on the right represents a different travel cost.

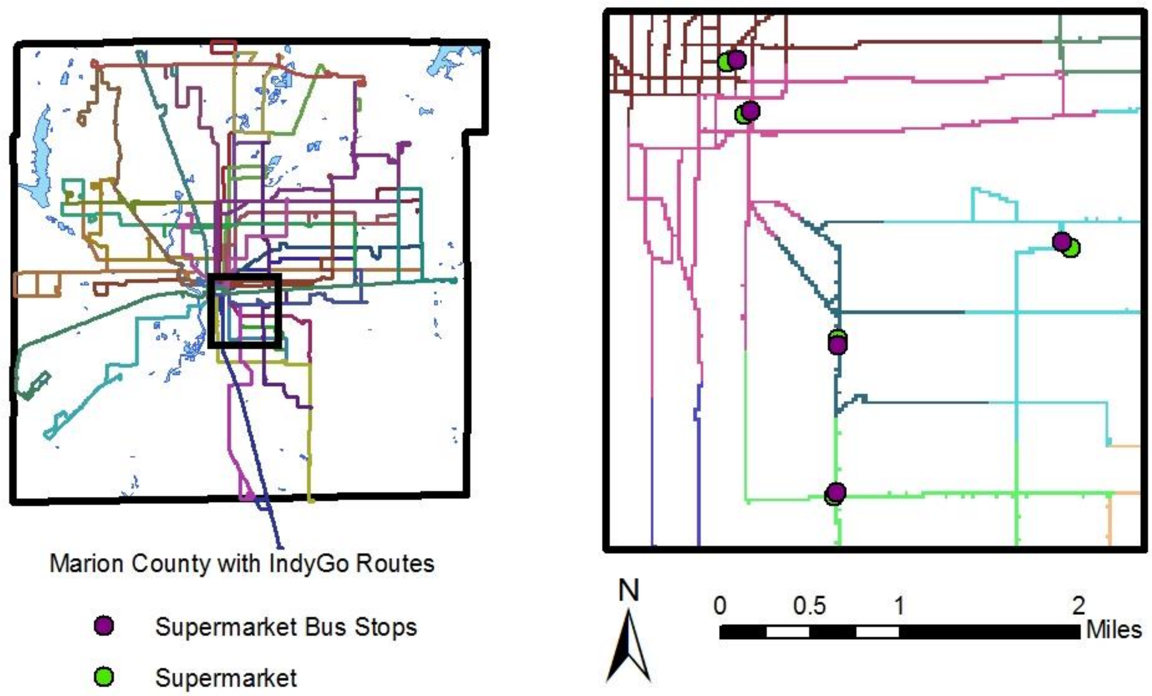

Figure 4-6: Allocation of Bus Network to Nearest Supermarket

Next, the *Cost Distance* tool is again used, this time with the transit network. Given the cost to traverse each cell, the lowest accumulative cost to the nearest supermarket bus stop is determined for every cell in the bus system. Lastly, the raster with the allocated costs from walking from the supermarket bus stops to the business completed earlier in this step is added to the cost distance raster and the bus stop raster, again subtracting a value of one for the bus stop raster's original value. Every bus stop now has a total cost representing the cost of travel time on the bus from that location to the supermarket bus stops and the cost of travel time from that supermarket bus stop to the nearest store.

4.3.4 Step 3: Distance from Each Cell to the Nearest Bus Stop

The first portion of the resident's trip is the final part of the travel cost determination. Similar to the previous step, the total cost at each bus stop must be converted to an integer using the *Int* tool in order to allocate the value to other cells. This integer cost value is then distributed to all of the cells in the walking network, giving each cell the total cost for the part of the journey after getting to the bus stop.

The *Cost Distance* tool is used for the walking network again, this time to find the least accumulative cost from each cell to the nearest bus stop. Finally, raster containing the cost of walking to the nearest bus stop is added to the raster representing the cost of the second two legs of the trip, and $1.75 is added to that total cost since IndyGo buses use that flat rate for any trip. Additionally, like the walking and driving analysis, the *Zonal Statistics* and *Zonal Statistics as Table* tools are used to aggregate the data for each census tract. Figures showing the total travel cost for each census tract can be seen in the following section.

4.4 Results

4.4.1 Travel Cost

The output from the *Zonal Statistics as Table* tool generates the one-way trip cost statistics, which are shown in Table 4-3. By only examining the travel cost, it is possible to identify areas that spend a higher amount on transportation to obtain healthy food. These census tracts can be seen in Figure 4-7, Figure 4-8, and Figure 4-9. Many are located in the periphery, or outside of I-465, and just north and northwest of the city center. The availability of transit is evident in the transit cost map, where the east-west and north-south corridors can be easily identified by the lower-cost census tracts.

Table 4-3: Descriptive Statistics for Travel Cost by Mode

	Walking	Driving	Transit
Mean	$3.04	$1.04	$4.12
Median	$2.74	$0.94	$3.05
Maximum	$13.30	$4.62	$18.55
Minimum	$0.84	$0.29	$2.10
Std. Dev.	$1.70	$0.57	$2.83

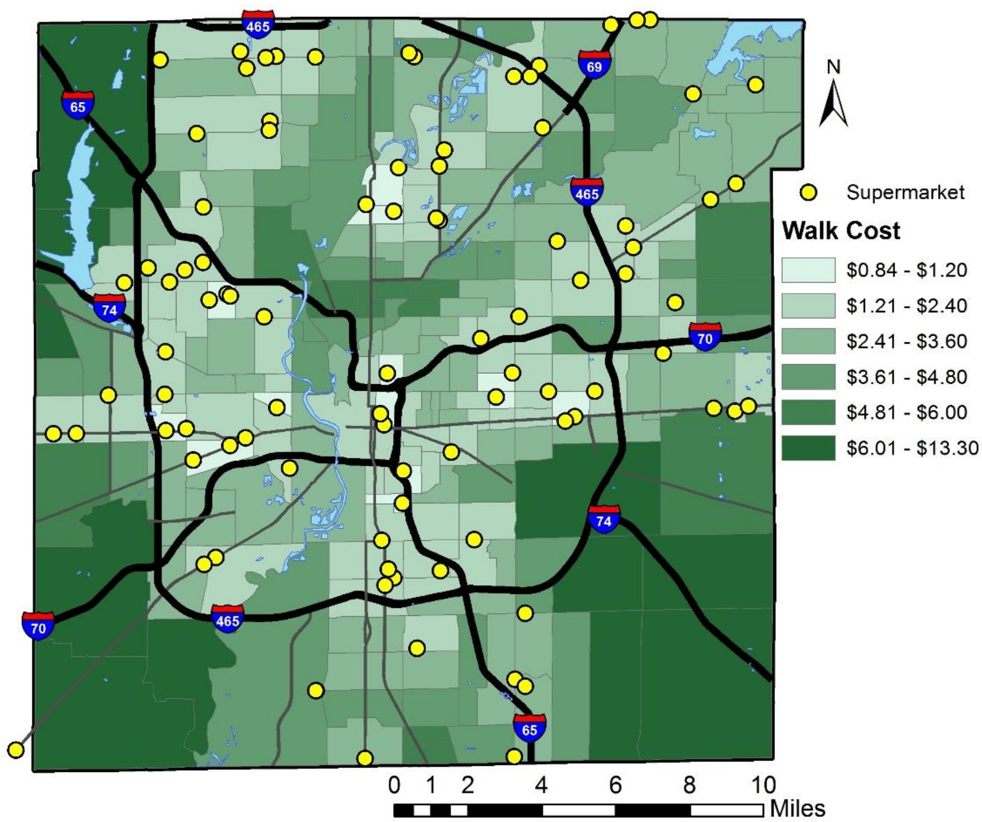

Figure 4-7: Average Travel Cost to the Nearest Supermarket by Walking

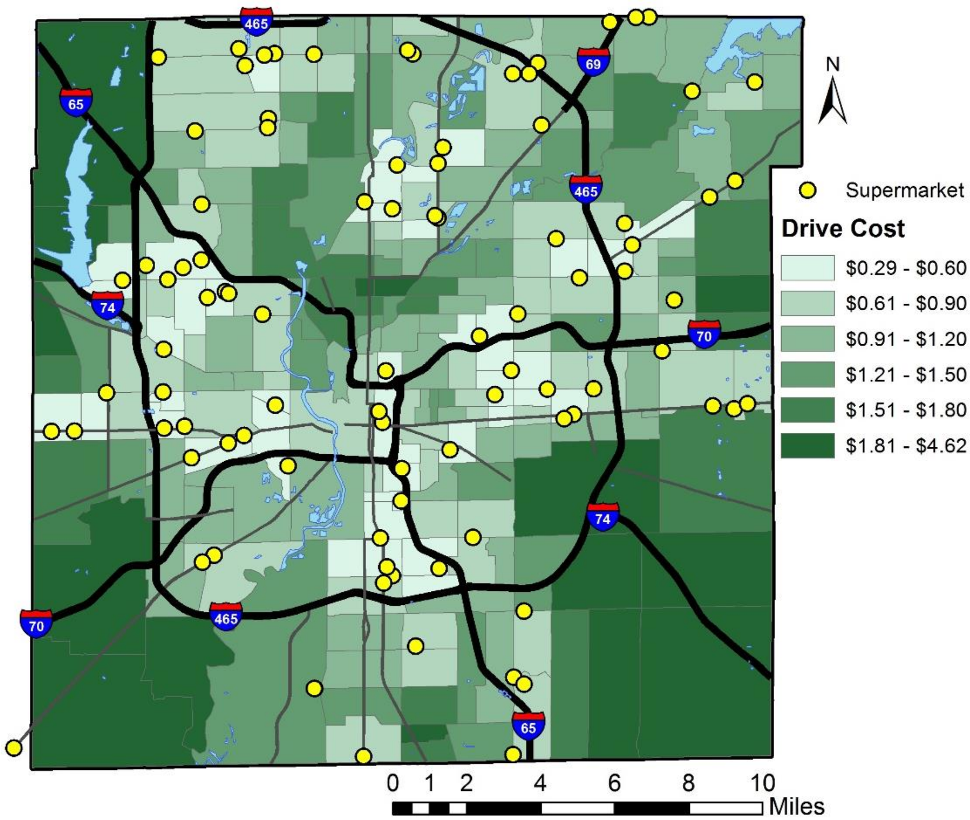

Figure 4-8: Average Travel Cost to the Nearest Supermarket by Driving

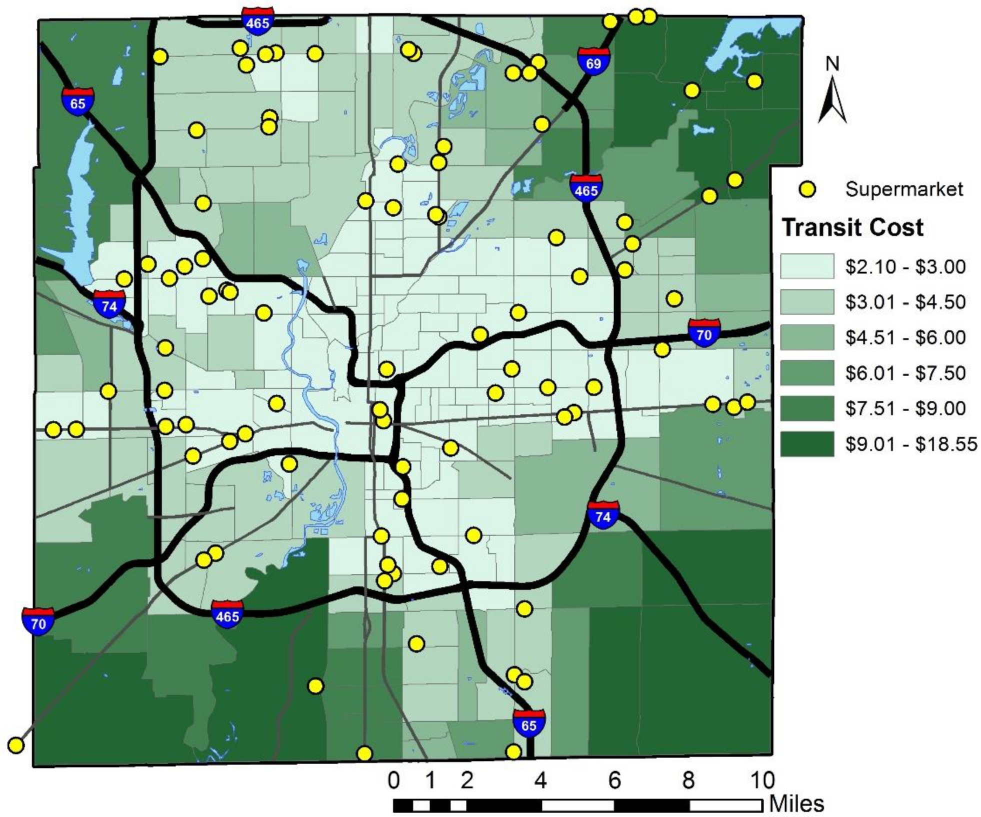

Figure 4-9: Average Travel Cost to the Nearest Supermarket by Transit

4.4.2 Travel Cost as a Proportion of Median Household Income

Although the above information indicates areas that have a higher travel cost, this does not necessarily imply that the area should be considered a food desert. It is important to consider income of the census tract to identify areas spending a disproportionate amount of their income, rather than simply a higher amount. First, one-way travel costs are converted to yearly travel costs by the following formula, assuming one trip per week to maintain a fresh stock of produce:

$$Y = cost * 2\frac{directions}{trip} * 1\frac{trip}{week} * 52\frac{weeks}{year} \qquad \text{Eq. 4-4}$$

These yearly travel costs are divided by the median household income; using yearly travel costs for this ratio makes the values much more understandable. Although a specific threshold to determine poor accessibility was not determined, the census tracts that have the highest ratios can be compared to the census tracts identified as food deserts by the USDA, as shown in Figure 4-10.

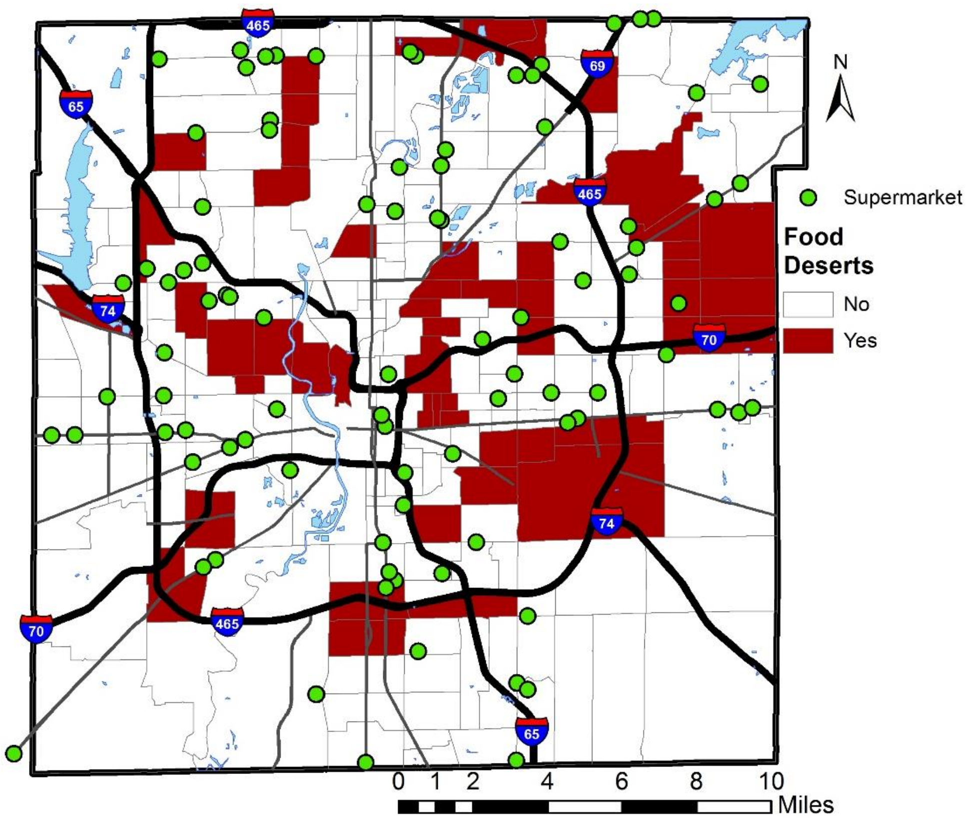

Figure 4-10: Census Tracts Designated as Food Deserts, adapted from (United States Department of Agriculture, 2013)

When compared with only the cost of walking or driving, the ratio as a measure of access eliminates many tracts in the periphery from consideration and emphasizes the lack of access in other areas, particularly the Near Eastside. The walking and driving ratios identify many of the same census tracts as the USDA as having poor access; however, certain areas show either less or more of a disadvantage. The Near Eastside, Near Westside, and tracts on the Far Eastside and in the southeastern portion of the county are consistently identified in both methods. However, areas of poor accessibility identified through the method outlined in this thesis include many tracts on the Near Northside and three tracts on the Far Southside near the county border. Areas that show improved accessibility when compared with USDA food deserts include several tracts in the northern third of the county. It is worth noting that land use should also be a consideration when determining potential areas for intervention. For instance, the tract showing the highest cost near the southwestern corner of the county is the location of the Indianapolis International Airport; zoning restrictions or Federal Aviation Administration requirements may affect the intervention's success, especially since the population of the tract is relatively low. The same can be said for census tracts that are highly industrial.

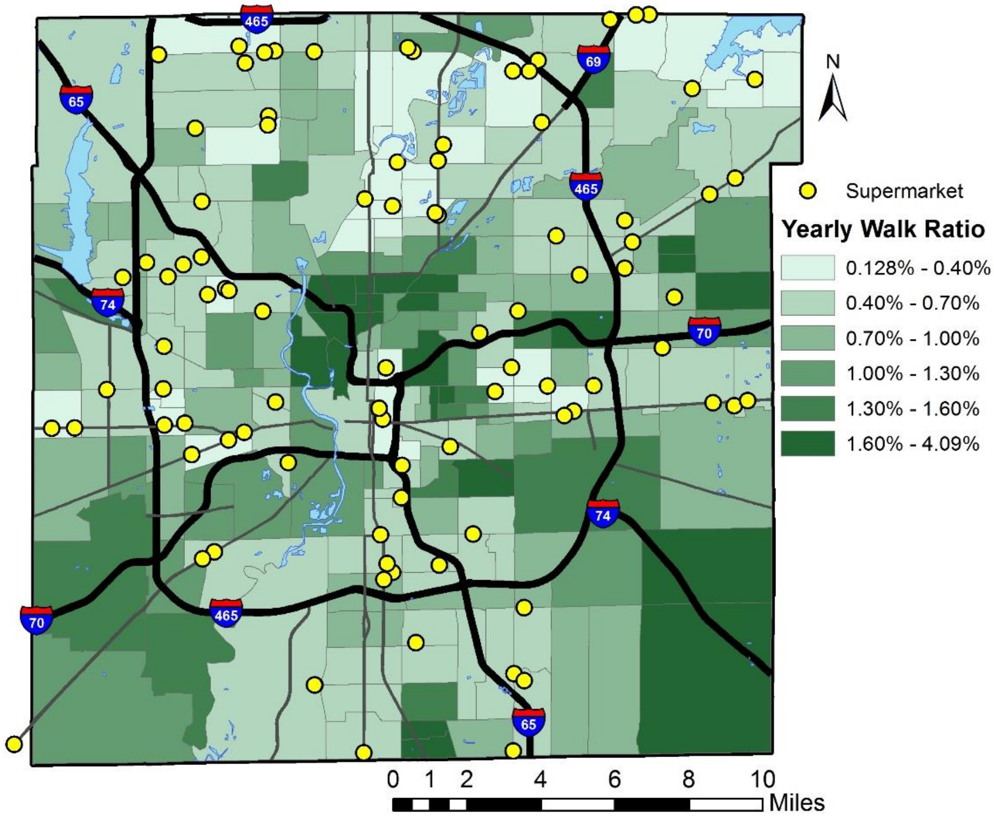

Figure 4-11: Percent of Income Spent on Walking to Supermarkets Annually

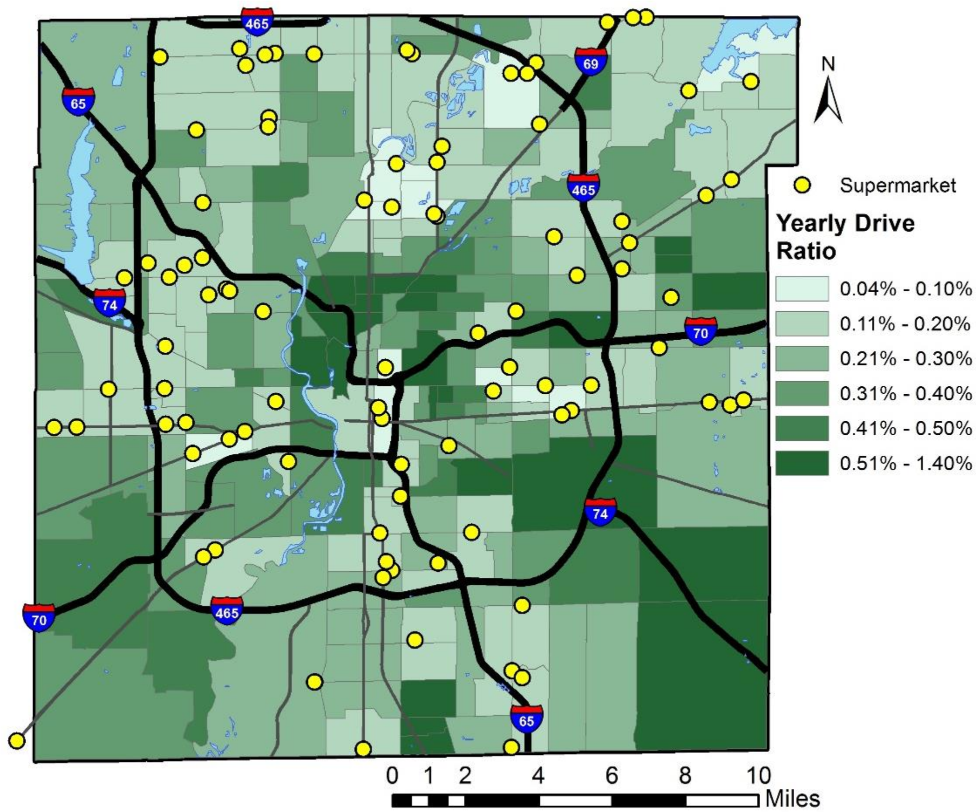

Figure 4-12: Percent of Income Spent on Driving to Supermarkets Annually

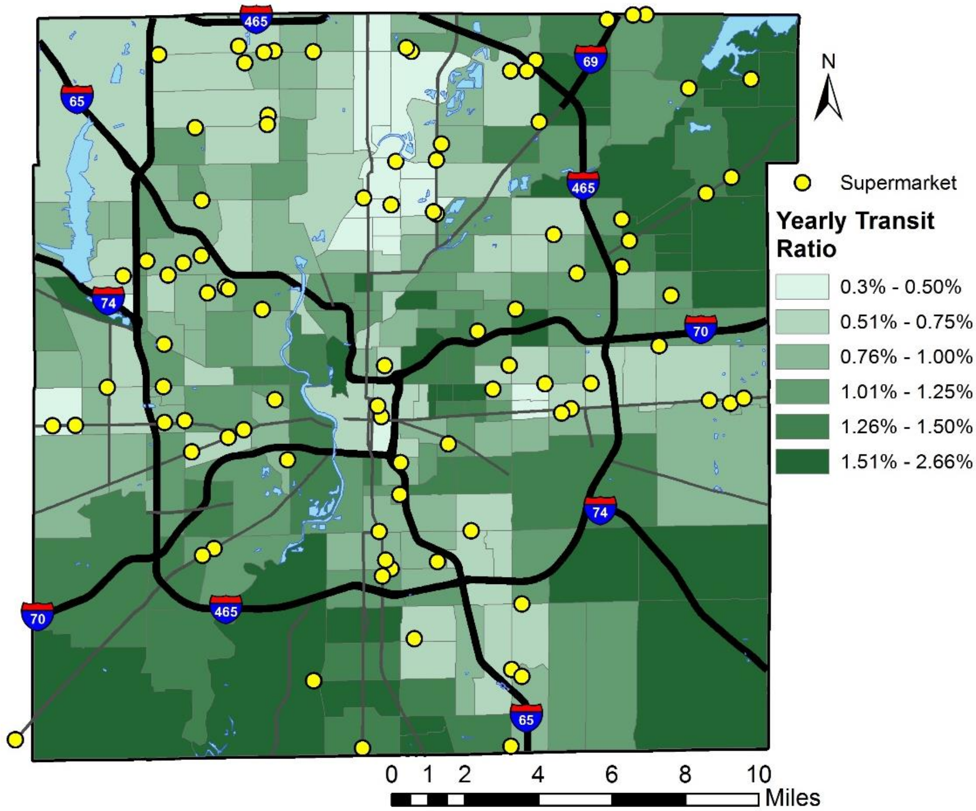

Figure 4-13: Percent of Income Spent on Taking Transit to Supermarkets Annually

Although the walking and driving analyses (Figure 4-11and Figure 4-12) have comparable implications since they consider very similar road networks, the transit analysis (Figure 4-13) has different implications. The higher ratios for tracts outside of the I-465 loop is largely indicative of the lack of transit service in those areas. The largest percentage of the trip cost would be to walk to the nearest bus stop, in which case residents would likely walk to the nearest supermarket instead. However, research indicates that consumers generally do not like to walk over 500 meters to obtain groceries (Wrigley, Warm, & Margetts, 2003); even at the higher end of that range, carrying heavy bags home could pose a problem, especially for the disabled or elderly.

Another important observation is that some areas have poor accessibility for all three modes. Most notably, the census tracts just east of downtown are known for being a low-income, high-crime area; these maps show that their residents are spending a much greater proportion of their income on transportation to get healthy food, regardless of which mode consumers choose. Other areas include the southeastern corner, the center along the southern county border, and just north of I-70 on the Far Eastside. Although the USDA method should not be discounted, analyzing food deserts by the available transportation network and as a function of median income could give a more accurate representation of areas that have poorer accessibility to healthy foods.

CHAPTER 5. STATISTICAL METHODOLOGY AND RESULTS

Areas that spend a higher proportion of their income on transportation to supermarkets were identified in the previous chapter; however, the analysis can be made more useful by determining any significant factors associated with poor accessibility. This chapter describes the statistical methods applied to identify correlations among cost, socioeconomic characteristics, and transportation factors. Significant variables, along with their marginal effects and elasticities, will be used as a basis for intervention recommendations.

5.1 Exploratory Regression

With a large number of potential explanatory variables, a starting point for the analysis was determined through the Exploratory Regression tool in ArcGIS. This tool allows for an input of up to twenty potential independent variables and identifies several combinations to best fit the dependent variable based on a specified number of minimum and maximum variables to be included. By allowing ArcGIS to construct best-fit models for different numbers of variables, some of the initial guess-and-check work to find an initial model is eliminated.

5.2 Ordinary Least Squares

After exploratory regression in ArcGIS, a model is chosen for each mode based on factors including the adjusted-R^2 value, number of variables, and the sign and significance of the variables. The R-squared value is a measure of fit calculated by the sum of square errors and the total variation of the dependent variable. However, since the R-squared value will increase for each variable added to the model (even if insignificant), the adjusted R-squared value also accounts for the number of variables in the model so as to not artificially inflate the goodness-of-fit measure (Washington, Karlaftis, & Mannering, 2011). The number of variables was also used as a selection criterion in order to simplify the model as much as possible. Two software programs are used to further analyze these models: GeoDa and GeoDaSpace. GeoDa is primarily used for exploratory spatial data analysis (ESDA) and maximum likelihood spatial regression; GeoDaSpace has more powerful spatial econometrics tools, some of which include allowing for different methods of estimating coefficients, the inclusion of endogenous variables, and non-normality or heteroscedasticity of error terms (GeoDa Center, 2015). The selected models from ArcGIS are consequently run in GeoDa and GeoDaSpace as ordinary least squares linear regression models. This model is of the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \qquad \text{Eq. 5-1}$$

where Y is the dependent variable, β_0 is the constant, β_n is the estimation of the unbiased coefficient, X_n is the independent variable for that observation, n is the number of explanatory variables included in the model, and ε is the error term. The dependent variable must be a continuous variable; in this case, the dependent variable is the natural log-transformed ratio of the average travel cost to the median household income. The error term also requires a number of assumptions: zero mean, homoscedasticity, no autocorrelation, no correlation with its observation, and normality (Washington et al., 2011). Tests for many of these assumptions will be discussed in the next section.

5.3 Diagnostics

5.3.1 Morans I

The Morans I statistic tests for spatial autocorrelation. Similar to the way in which the Durbin-Watson statistic can indicate serial autocorrelation in time-series data (i.e., over time), this statistic captures any relationship between the regression residuals and their spatially lagged values (i.e., over space) (Arbia, 2014). Although there is no explicit alternative hypothesis, the statistic tests the probability that the errors exhibit spatial autocorrelation.

5.3.2 Normality

The Jarque-Bera statistic tests the normality of the residuals. The statistic considers the symmetry and kurtosis and is χ^2-distributed with two degrees of freedom. If the statistic is significant, it could indicate a misspecified model with potentially significant missing variables.

5.3.3 Heteroscedasticity

The Breusch-Pagan statistic is a measure of heteroscedasticity of the errors, which indicates that as the size of the regressor increases, the size of the disturbance also increases. It tests the null hypothesis that the disturbance variance is constant and is χ^2-distributed with degrees of freedom equal to the number of variables plus 1. If this value is significant, the errors likely exhibit heteroscedasticity, but this can be solved by using a generalized method of moments (GMM) with heteroscedasticity in GeoDaSpace. GMM uses an initial estimate of the spatial autoregressive parameter λ to predict the coefficients through spatially weighted least squares regression. Unlike the maximum likelihood model, which uses a constant error variance as one of its conditions, equations to solve the GMM model do not assume a constant error variance to allow for heteroscedasticity (Anselin & Rey, 2014). However, some heteroscedasticity is caused simply by the process of spatial dependence (Florax, 2015).

5.4 Spatial Regression

In some cases, spatial autocorrelation in the dependent values or the errors can be improved by using a spatial regression model. The most prominent spatial regression models are the spatial lag model and the spatial error model, which are commonly used to model information with spatial spillovers or effects. For example, (Chen, Florax, Snyder, & Miller, 2010) used a spatial lag model to examine the relationship between body mass index and characteristics of the surrounding communities. Additionally, a spatial error model was used to find any socioeconomic factors that were correlated with better access to supermarkets for transportation analysis zones when considering their transit commute patterns (Widener et al., 2015).

5.4.1 Weights Matrix

The weights matrix W is the differentiating factor between spatial regression models and ordinary least squares regression. The row-standardized matrix is $n \times n$ and denotes neighbors of the spatial unit. The matrix can be defined using either rook contiguity, which considers neighbors as any adjacent unit with a shared border, or queen contiguity, which considers neighbors as those units with shared borders and shared vertices (Anselin & Rey, 2014). The order of contiguity must also be specified, or the extent of the neighbors considered. For example, a first-order contiguity matrix would only include direct neighbors, but a second-order contiguity matrix would include weights for the neighbors' neighbors as well. In this research, a first-order queen contiguity matrix is used.

5.4.2 Model Specification

Once the weights matrix is defined, the spatial regression models can be explored. The proper spatial model can be selected using the Lagrange Multiplier statistics specified in the following section. This discussion is adapted from Anselin and Rey, 2014.

The spatial lag model uses the values of the neighbors and an autoregressive parameter ρ to better predict the dependent variable by accounting for spillover effects. This model is of the form

$$y = \rho W y + X\beta + u \qquad \text{Eq. 5-2}$$

where y is the dependent variable, ρ is an autoregressive coefficient, W is the spatial weights matrix, X is a vector of the regressors, β is a vector of the estimated coefficients, and u is the error vector. Because of the spatially lagged dependent variable, this equation is commonly written in its reduced form, where $(I - \rho W)^{-1}$ is considered a spatial multiplier.

$$(I - \rho W)y = X\beta + u \qquad \text{Eq. 5-3}$$

$$y = (I - \rho W)^{-1}X\beta + (I - \rho W)^{-1}u \qquad \text{Eq. 5-4}$$

Similar to the way in which the error term in an OLS model captures some correlation among regressors and effects of omitted variables, the spatial error model accounts for omitted variables that exhibit spatial correlation and is specified as

$$y = X\beta + u \text{ and } u = \lambda W u + \varepsilon \qquad \text{Eq. 5-5}$$

where u is the error vector. The error vector can be rewritten as $u(1 - \lambda W) = \varepsilon$, so in order to remove the endogeneity from the error vector, each side of the equation can be multiplied by $(I - \lambda W)^{-1}$. This multiplier is considered a spatial filter and gives a reduced form of $u = \varepsilon(1 - \lambda W)^{-1}$. Therefore, the reduced form of the spatial error model is

$$y = X\beta + \varepsilon(I - \lambda W)^{-1} \qquad \text{Eq. 5-6}$$

Once other variables were added and checked for significance, the new best model was run in OLS again to check the Lagrange Multiplier statistics and ensure the model was still properly specified.

5.4.3 Lagrange Multipliers

In order to properly specify the model, Lagrange Multiplier (LM) statistics and Robust Lagrange Multiplier (RLM) statistics are calculated for the spatial lag model and the spatial error model. The Lagrange Multiplier tests require inputs from the ordinary least squares regression model, which must be determined first, and test for spatial dependence. The null hypotheses are that $\lambda = 0$ and $\rho = 0$ for the error and lag model, respectively. If the null hypothesis for each model type fails to be rejected, then an OLS model is likely appropriate since neither of the spatial autoregressive parameters can be said to be not equal to zero. The statistic follows a χ^2 distribution with one degree of freedom. If the spatial lag LM statistic is significant, the spatial lag model should be chosen; if the spatial error LM statistic is significant, the spatial error model should be chosen. However, it is possible for both the spatial lag and spatial error statistics to be significant. In that case, the RLM statistics can be used. Similar to the process for the LM statistics, the more significant value indicates which model should be chosen. If both are still significant at high significance levels, it could indicate that the model is misspecified.

5.5 Model Fit

In order to select the best-fit model, certain characteristics of model fit must be selected for analysis. First is the spatial pseudo-R^2, which is similar to the standard R^2 calculated for ordinary least squares models. It is a measure of correlation between the predicted values using the reduced form (Equation 5-4) and observed predicted variables. The log-likelihood value was also considered; model development for OLS and Maximum Likelihood methods entails maximizing the log-likelihood, so a higher number indicates a better fit (Anselin & Rey, 2014). Lastly, the condition number indicates the degree of multicollinearity in the model, so a lower value is more ideal. That said, each of these measures does not have an explicit hypothesis to say whether or not the model is a good fit; they simply serve as comparative statistics.

5.6 Marginal Effects

Marginal effects can be interpreted as the effect on the dependent variable of a one-unit change in the explanatory variable. They are calculated as the derivative of the dependent variable with respect to the explanatory variable of interest and most frequently used to describe the effects of count, discrete, or indicator variables. Anselin and Rey (2014) derive the marginal effect equation for a linear spatial lag model, where the first term is interpreted as the direct effect and the second term uses the spatial multiplier to account for indirect effects (Equation 5.7). Since this equation contains values for the direct and spillover effects, this is interpreted as the global effect of a one unit change in each spatial unit.

$$\frac{\partial y}{\partial x} = \beta_h + \frac{\beta_h \rho}{1-\rho} = \frac{\beta_h}{1-\rho} \qquad \text{Eq. 5-7}$$

Although Equation 5.7 is a good estimate of the overall effects, the effects can be calculated directly in R by multiplying the inverse of the spatial multiplier matrix by the estimated coefficient and then taking the average of the row sums, which account for the direct and indirect effects. For this study, the analysis of the effects must also account for the log-transformed dependent variable. For log-linear models, unit changes in independent variables are interpreted as percentage changes in the dependent variable (Wooldridge, 2002). The percentage change in y is given by the following formula, which holds true for both continuous and dummy variables:

$$\%\Delta y = 100(e^{\beta \Delta x} - 1) \qquad \text{Eq. 5-8}$$

Then, given the results from the R output and assuming that $\Delta x = 1$, Equation 5.8 can be derived to include the spatial multiplier to represent the total effect on the dependent variable:

$$\%\Delta y = 100(e^{\frac{\beta_h}{1-\rho}} - 1) \qquad \text{Eq. 5-9}$$

5.7 Results

5.7.1 Model Significance

Although many different variables were collected from several different sources, only three of those variables were determined to be significant. No data from the NHTS was used; average vehicles per household was originally found to be significant but removed because of endogeneity concerns. The models developed used the same three significant variables. An education variable, NOCOLLEGE, represents the number of people over the age of 25 without any college degree. Another variable, CRIMEDEN, stands for the crime density of the census tract, which is calculated by the number of crimes in that census tract divided by its area in square miles. The third variable, INLOOP, is an indicator variable distinguishing the urbanized center from the periphery. In this case, a census tract was selected as being in the urban center if a majority of its area was inside the I-465 beltway. This variable accounts for some of the heterogeneity in the data. Descriptive statistics of these variables can be found in Table 5-1. Although GeoDaSpace interprets missing values as zero, the mean, standard deviation, and minimum shown here are exclusive of those missing values.

Table 5-1: Descriptive Statistics of Significant Variables

Variable	***Mean or %***	***Std. Dev.***	***Obs.***	***Maximum***	***Minimum***
NOCOLLEGE	1862.48	809.86	223	5608	5
CRIMEDEN	100.67	110.59	218	657.66	0.81
INLOOP	68.30%	--	224	--	--

These coefficients were analyzed using two-tailed hypothesis testing. The statistic follows a normal distribution and were tested at a significance level of 10%. The null hypothesis and confidence level equations are written below. Additionally, the results from the spatial lag model regression analysis, including significant variables and measures of fit, are shown in Table 5-2.

$$H_0\colon \beta = 0, \qquad z \geq |1.645|$$

Table 5-2: Model Specifications for Each Mode (t-statistics in parentheses)

Variable	***Walking***	***Driving***	***Transit***
Constant	0.4027 (3.0219)	0.0582 (0.6502)	0.6274 (4.3641)
INLOOP	-0.1163 (-1.7371)	-0.1110 (-1.7009)	-0.1592 (-3.2370)
NOCOLLEGE	0.00009 (2.5587)	0.00009 (2.5386)	0.00005 (1.9269)
CRIMEDEN	0.00103 (3.5443)	0.00102 (3.5906)	0.00066 (3.2687)
Model Statistics			
ρ	0.6872	0.7018	0.6936
Spatial pseudo-R^2	0.1957	0.1935	0.2513
Log-likelihood	-141.800	-136.932	-62.488
Condition Number	6.503	6.503	6.503

Although the constant in the driving equation is not significant, it will be included in the model to account for some unobserved effects. Therefore, using the values from the table above, the equations can be written as follows, where access is defined as the travel cost divided by the median household income of the tract and u is the error term.

$$walking\text{:}\ (\mathbf{I}-0.6872\mathbf{W})\times\ln(access) = 0.4027-0.1163*INLOOP+0.00009*NOCOLLEGE+0.00103*CRIMEDEN+\boldsymbol{u_w} \qquad \text{Eq. 5-12}$$

$$driving\text{:}\ (\mathbf{I}-0.7018\mathbf{W})\times\ln(access) = 0.0582-0.1110*INLOOP+0.00009*NOCOLLEGE+0.00102*CRIMEDEN+\boldsymbol{u_d} \qquad \text{Eq. 5-13}$$

$$transit\text{:}\ (\mathbf{I}-0.6936\mathbf{W})\times\ln(access) = 0.6274-0.1592*INLOOP+0.00005*NOCOLLEGE+0.00066*CRIMEDEN+\boldsymbol{u_t} \qquad \text{Eq. 5-14}$$

5.7.2 Marginal Effects

The values for marginal effects can be seen in Table 5-3. After multiplying the coefficients by the spatial multiplier, the effects are calculated using Equation 5.8. Because of the logarithmic form, the results are given as percentage increases or decreases. It can be noted that the effects of these factors on the walking and driving modes are very similar. These effects show that living inside the I-465 loop, a more urbanized area as opposed to suburban neighborhoods, results in a decrease in proportion of income spent of approximately 37.2% for walking and driving and 51.94% for transit. An increase of one additional crime per square mile in each census tract results in a percent increase in income spent of 0.33% for walking and 0.34% for driving but only 0.22% for transit. Similarly, an increase of one additional person without a college degree results in an increase of approximately 0.029% for walking and driving but only 0.0156% for transit. These marginal effects indicate that residents inside the I-465 beltway spend a significantly smaller proportion of their income on transportation to obtain groceries, but costs for residents that use transit are less sensitive to changes in education or crime.

Table 5-3: Marginal Effects of Significant Variables by Mode

Variable	***Walking***	***Driving***	***Transit***
NOCOLLEGE	0.0290%	0.0294%	0.0156%
CRIMEDEN	0.3300%	0.3426%	0.2154%
INLOOP	-37.17%	-37.22	-51.94

5.7.3 Implication of Findings

These significant factors could be linked to intervention strategies that have been identified by previous research. Several studies have examined the addition of food outlets, but first it is important to identify whether the lack of a supermarket is a due to a shortage of demand or supply. Interventions may not be successful if they do not have an effect on the cause of the problem. Bitler and Haider (2011) explored food deserts from an economic perspective and suggested attempting to determine whether the food desert is caused by demand or supply factors. For instance, supply-side government interventions may help a food desert caused by the unwillingness of supermarkets to locate in that area; on the other hand, demand-side problems like low income or a lack of education about healthy foods would be more alleviated by interventions such as increased SNAP benefits or education programs.

Applicable aspects of their research for this study suggest that crime may be a supply-side problem and lack of education could be a demand-side problem. Intervention effects can be explicitly modeled in order to choose the best type and location. Because of the spatial multiplier used in this model, changes in explanatory variables that are not constant over the study area yield different marginal effects for each spatial unit. Programs such as R can be used to model the effects of specific interventions for each unit, allowing more individual effects to be seen across the study area. Since the crime density values exhibit obvious spatial clustering (see Figure 3-2), an intervention for this characteristic and its effects on the walking measure of access was chosen as an example. A 1% reduction in crime density was calculated for only the five tracts with the highest crime density, four of which are on the Near Eastside and one of which is on the Near Northside. Both of the other variables as well as the crime density in all other tracts were held constant. Using the coefficients and spatial multipliers that were already estimated, the effects in those tracts can be estimated by matrix multiplication. Percentage changes in the walking measure of access for each census tract are shown in Figure 5-1. A 1% decrease in crime in just those five tracts has an average effect of -0.0365% on the walking measure of access over the entire county.

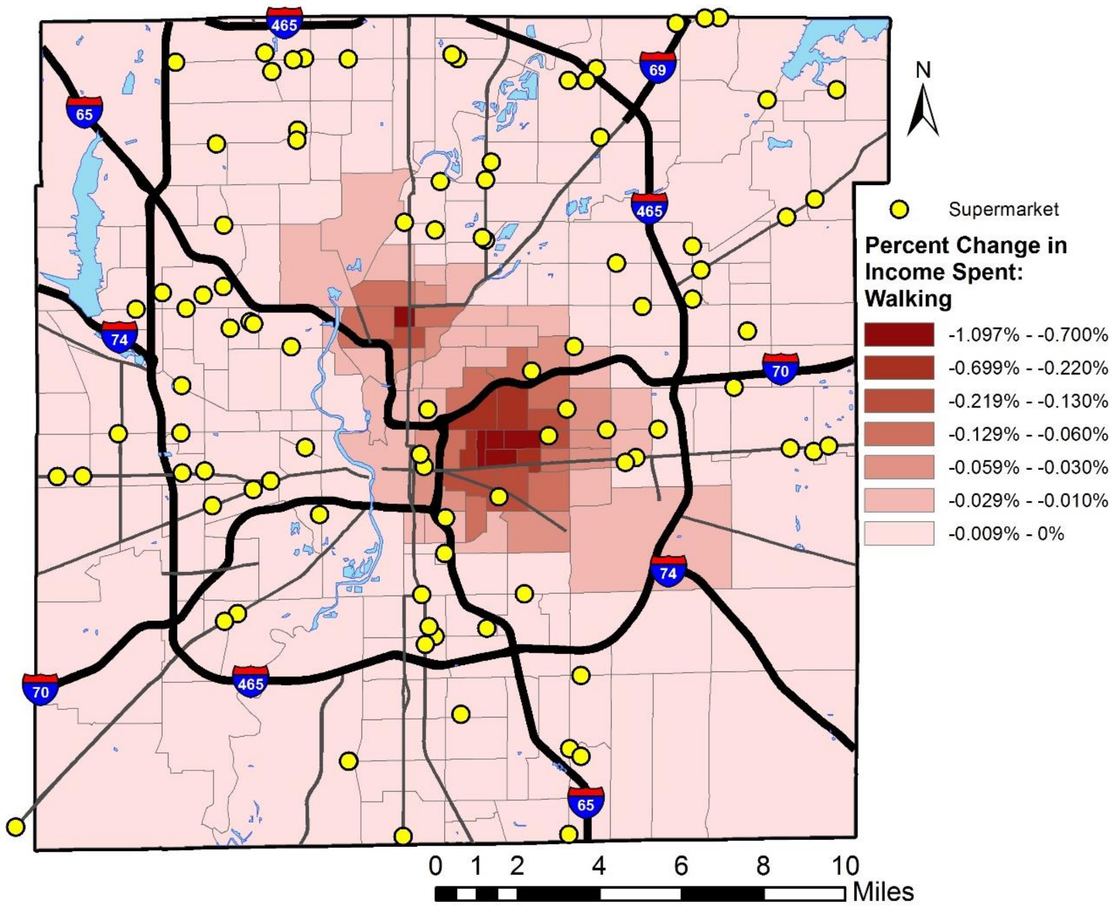

Figure 5-1: Marginal Effects of a Crime Intervention on the Walking Proportion of Income

These areas chosen for intervention are all consistent with census tracts designated as having poor accessibility to healthy food. Reducing crimes in these neighborhoods may attract grocery stores to the area due to improved neighborhood safety and a lower risk of property crimes such as theft. This is consistent with the Quality of Life plan for the Near Eastside, which states improved safety as one of their priorities through a "renewed partnership with the Indianapolis Metropolitan Police Department and an increase in block clubs" (Local Initiatives Support Corporation Indianapolis, 2012).

Another potential intervention is the expansion of the transit system. A noticeable lack of transit system outside I-465 contributes to the higher costs for census tracts in the periphery, and the marginal effect of that variable in the transit model is greater than that of either walking or driving. Although research in Cincinnati found that only two transportation analysis zones (TAZs) showed improved access when considering commuting patterns by automobile (Widener et al., 2013), zones with at least 10 transit commuters saw improvements in supermarket access when considering transit commutes in 28 out of 233 TAZs (Widener et al., 2015).

As mentioned previously, these implications all assume one grocery trip per week in order to maintain a stock of fresh produce. Areas spending a larger proportion of their income on transportation may choose to shop for groceries less frequently in order to save money, leading to less fruit and vegetable consumption and possibly increasing obesity rates. Widener, Metcalf, & Bar-Yam (2013) estimated the locations of low-income households and modeled the effects of changing households from biweekly or monthly grocery shopping trips to weekly trips. They found that increasing the shopping frequency of half of the households that shopped biweekly or monthly caused the number of households with fresh produce to increase by more than 8%, even more effective than the addition of a mobile market.

CHAPTER 6. CONCLUSIONS

6.1 Summary and Contribution of this Thesis

Referring back to the research objectives, this thesis had two primary purposes: to develop a measure to accurately determine areas with poor supermarket accessibility, and to determine the extent to which the socioeconomic and transportation environments affect the level of accessibility. To achieve these objectives, the author used data for Marion County, Indiana in 2012.

In order to accomplish the first objective, Spatial Analyst in ArcGIS was used to calculate travel cost to the nearest supermarket by creating rasters in which the cells are associated with a specific cost for each mode. It expands on previous methodology by considering additional information, such as the cost of operating a motor vehicle in addition to the value of time. It also considers the three trip segments necessary for transit—walking to the bus stop, riding the bus, and walking from the bus stop to the supermarket—through a combination of cost distance and cost allocation processes. A ratio of travel cost to median household income is then used to define areas with poor accessibility to healthy food. Previously, the income and access components were considered in conjunction, but separately, excluding areas from consideration that may fall near but not below one of the thresholds. This method enables the comparison of census tracts identified as spending the highest proportion of income with those that are designated food deserts by the USDA in

2015. Although many tracts are consistently identified, others are either added or eliminated from consideration.

The second task was to identify any effect of the environment on this measure of accessibility. Given the proportion of income spent on transportation to supermarkets for each census tract, a spatial lag model was estimated to identify any correlational socioeconomic factors. For each mode, the same three significant variables were found: crime density and the number of people without a college degree were found to be directly related to the ratio of travel cost to median household income, and living inside I-465 was found to decrease that ratio. Marginal effects were calculated in order to compare mode sensitivity and identify potential areas for interventions. Walking and driving had similar effects, but transit access to supermarkets was less sensitive to changes in crime and education and more sensitive to location. Potential interventions that were identified included expanding transit service to more areas outside of I-465 and taking measures to reduce crime in neighborhoods like the Near Northside and particularly the Near Eastside. These interventions could help give residents better access to supermarkets, thereby increasing availability of healthy food and potentially decreasing obesity rates.

This thesis has contributions within this field of study as well as extensions in other fields. First, the travel cost is more accurately calculated. The driving cost includes both a value of time and an operating cost, and the transit analysis considers all three trip segments as well as a flat bus fare. Additionally, this thesis defines census tracts with poor supermarket accessibility as a ratio of travel cost to median household income. Outside this field of study, this information could be used by local metropolitan planning organizations (MPOs) or other government agencies. In this instance, the Indianapolis MPO and the

Indiana Economic Development Corporation could benefit from this data. Purdue Extension offices, which work to educate the community in the areas of agriculture, health and human sciences, and community development, could use this data as well.

6.2 Limitations

Limitations in this thesis primarily stem from a lack of data, which required assumptions to be made. Additional data pertaining to the roads could eliminate a few of these assumptions. First, speed limits had to be generalized by road type because speed limits for specific road segments were not available; having this information would increase the accuracy of the travel cost estimate. Further, knowing whether or not the road segment had a sidewalk could improve the walking analysis by excluding certain segments without sidewalks or by considering safety. The availability of socioeconomic data also poses certain limitations. The Indianapolis Metropolitan Police Department does not have jurisdiction over some small incorporated areas within the county, so crime information from local jurisdictions was necessary; however, the level of detail is significantly less than that of the IMPD and in one instance unavailable entirely. The analysis could also potentially be improved by disaggregating its spatial level, but household income is only available at the census tract level and not at the block level for Marion County.

Other assumptions were required in order to simplify the analysis in ArcGIS. First, it was assumed that consumers would choose the supermarket closest to their residence in order to use the *Cost Distance* tool; however, due to cost constraints or other preferences, this may not be the consumer's supermarket of choice. Without considering preferences, this method conservatively estimates a resident's cost of healthy food access, but cost constraints could negatively affect consumers if they are unable to afford the nearest supermarket. Other aspects not considered were frequency of transit and the added time and cost of changing bus lines. Accounting for a trip on two separate routes was beyond the scope of this research, and it was assumed that the frequency would be partially taken into account by the consumer in determining when to leave the house and how long to spend at the store.

6.3 Recommendations for Future Research

In addition to expanding upon the limitations listed above, other recommendations for expanding research on this topic are discussed briefly in this section. Although many costs are tangible, other perceived costs can impact consumers' decisions. Quantifying aspects such as safety, whether in high-crime areas or on roads without sidewalks, may more accurately reflect the travel cost perceived by residents of that neighborhood. It could also give insight to the probability that consumers will choose a particular mode by estimating a discrete choice model. From this analysis, other interventions like adding sidewalks to a busy street could be explored if the related variable was found to have a significant effect, or it could identify the mode for which an intervention would be most effective. Along the same lines, because the travel costs are typically related to an extent due to the existing transportation infrastructure, accessibility measures could be modeled using a system of equations as well; this could allow differences among modes to be better seen. However, currently, spatial regression models are unable to account for multiple equations.

Another potential topic for future research would be to investigate changes in supermarket availability over a certain time period. If modeled against socioeconomic factors and transportation network availability over time, elements that correlate with the existence of supermarkets in neighborhoods could be found in order to recommend intervention strategies with the best chance of success. Additionally, the effect of changes in supermarket availability over time on obesity rate trends could be examined. A strength of a longitudinal rather than cross-sectional study in this research area is that the population would remain largely the same, and changes are less likely to be attributed to unobserved characteristics.

REFERENCES

Anselin, L., & Rey, S. J. (2014). *Modern Spatial Econometrics in Practice.* Chicago, IL: GeoDa Press LLC.

Arbia, G. (2014). *A Primer for Spatial Econometrics.* New York: Palgrave MacMillan.

Bader, M., Perciel, M., Yousefzadeh, P., & Neckerman, K. (2010). Disparities in neighborhood food environments: Implications of measurement strategies. *Economic Geography, 86*(4), 409-430.

Bitler, M., & Haider, S. J. (2011). An economic view of food deserts in the United States. *Journal of Policy Analysis and Management, 30*(1), 153-176. doi:10.1002/pam.20550

Bodor, J. N., Rice, J. C., Farley, T. A., Swalm, C. M., & Rose, D. (2010). Disparities in food access: Does aggregate availability of key foods from other stores offset the relative lack of supermarkets in African-American neighborhoods? *Preventative Medicine, 51*, 63-37. doi:10.1016/j.ypmed.2010.04.009

Breyer, B., & Voss-Andreae, A. (2013). Food mirages: Geographic and economic barriers to healthful food access in Portland, Oregon. *Health & Place, 24*, 131-139.

Bureau of Transportation Statistics. (2014, September). *Table 3-17: Average Cost of Owning and Operating an Automobile*. Retrieved from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_03_17.html

Burns, C. M., & Inglis, A. D. (2007). Measuring food access in Melbourne: Access to healthy and fast foods by car and foot in an urban municipality in Melbourne. *Health & Place, 13*, 877-885. doi:doi:10.1016/j.healthplace.2007.02.005

Centers for Disease Control and Prevention. (2012). *Behavioral Risk Factor Surveillance System: Prevalence and Trends Data*. Retrieved from http://apps.nccd.cdc.gov/brfss/

Centers for Disease Control and Prevention. (2012). *Sortable Risk Factors and Health Indicators*. Retrieved from http://wwwn.cdc.gov/sortablestats/

Centers for Disease Control and Prevention. (2013, May 9). *SMART: BRFSS City and County Data*. Retrieved from http://apps.nccd.cdc.gov/BRFSS-SMART/

Centers for Disease Control and Prevention. (2015, February 23). *Assessing Your Weight: About BMI for Adults*. Retrieved from http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/

Chen, S., Florax, R. J., Snyder, S., & Miller, C. C. (2010). Obesity and Access to Chain Grocers. *Economic Geography, 86*(4), 431-452.

Coveney, J., & O'Dwyer, L. (2009). Effects of mobility and location on food access. *Health & Place, 15*, 45-55.

Eckert, J., & Shetty, S. (2011). Food systems, planning, and quantifying access: Using GIS to plan for food retail. *Applied Geography, 31*, 1216-1223. doi:doi:10.1016/j.apgeog.2011.01.011

Federal Bureau of Investigation. (2012). *Uniform Crime Reporting Statistics*. Retrieved from http://www.ucrdatatool.gov/Search/Crime/Local/OneYearofDataStepTwo.cfm

Federal Highway Administration. (2011). *Introduction to the 2009 NHTS*. Retrieved from http://nhts.ornl.gov/introduction.shtml

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributed to obesity: Payer- and service-specific estimates. *Health Affairs, 28*(5), 822-831. doi:10.1377/hlthaff.28.5.w822

Florax, R. J. (2015, February 1). Spatial Regression Models. West Lafayette, IN: Department of Agricultural Economics, Purdue University.

GeoDa Center. (2015). *GeoDa Center and Affiliated Software*. Retrieved from https://geodacenter.asu.edu/software

Google Developers. (2015, February 2). *General Transit Specification Feed Reference*. Retrieved from https://developers.google.com/transit/gtfs/reference

Gordon, C., Purciel-Hill, M., Ghai, N. R., Kaufman, L., Graham, R., & Van Wye, G. (2011). Measuring food deserts in New York City's low-income neighborhoods. *Health & Place, 17*, 696-700. doi:10.1016/j.healthplace.2010.12.012

Guiliano, G. (2005). Low income, public transit, and mobility. *Transportation Research Record: Journal of the Transportation Research Board, 1927*, 63-70.

Hallett, L. F., & McDermott, D. (2011). Quantifying the extent and cost of food deserts in Lawrence, Kansas, USA. *Applied Geography, 51*, 1210-1215.

Horner, M. W., & Wood, B. S. (2014). Capturing individuals' food environments using flexible space-time accessibility measures. *Applied Geography, 51*, 99-107.

Indianapolis Metropolitan Police Department. (2010). *IMPD Jurisdiction*. Retrieved from http://www.indy.gov/CRIMEWATCH/IMPDDISTRICTS/Pages/IMPDDistricts.aspx

Indianapolis Metropolitan Police Department. (2012). *UCR Downloads*. Retrieved from http://www.indy.gov/egov/city/dps/impd/crimes/pages/ucrdownload.aspx

IndyGo. (2014, June 13). *Developer Content*. Retrieved from http://www.indygo.net/about-indygo/indygo-developer-content-area/

James, P., Arcaya, M. C., Parker, D. M., Tucker-Seeley, R. D., & Subramanian, S. (2014). Do minority and poor neighborhoods have higher access to fast-food restaurants in the United States? *Health & Place, 29*, 10-17. doi:http://dx.doi.org/10.1016/j.healthplace.2014.04.011

Local Initiatives Support Corporation Indianapolis. (2012). *Near Eastside Quality of Life Plan.* Indianapolis: LISC Institute for Comprehensive Community Development.

Morang, M. (2014). *Using GTFS Data in ArcGIS Network Analyst*. Retrieved from http://transit.melindamorang.com/

Smoyer-Tomic, K. E., Spence, J. C., & Amrhein, C. (2006). Food deserts in the prairies? Supermarket accessibility and neighborhood need in Edmonton, Canada. *The Professional Geographer, 58*(3), 307-326.

Sohi, I., Bell, B. A., Liu, J., Battersby, S. E., & Liese, A. (2014). Differences in food environment perceptions and spatial attributes of food shopping between residents of low and high food access areas. *Journal of Nutrition Education and Behavior, 46*(4), 241-249.

Tomer, A., Kneebone, E., Puentes, R., & Berube, A. (2011). *A Missed Opportunity: Transit and Jobs in Metropolitan America.* Washington, D.C.: Metropolitan Policy Program, The Brookings Institution.

United States Census Bureau. (2013). *Table DP03: Selected Economic Characteristics*. Retrieved from http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

United States Census Bureau. (2013). *Table S0802: Means of Transportation to Work by Selected Characteristics.* . Retrieved from http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

United States Census Bureau. (2015, February 9). *MAF/TIGER Feature Class Code Definitions*. Retrieved from https://www.census.gov/geo/reference/mtfcc.html

United States Department of Agriculture. (2013, May 8). *Food Access Research Atlas*. Retrieved from http://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data.aspx

USDA Agricultural Marketing Service. (n.d.). *Creating Access to Healthy, Affordable Food*. Retrieved from http://apps.ams.usda.gov/fooddeserts/fooddeserts.aspx

Ver Ploeg, M., Breneman, V., Dutko, P., Williams, R., Snyder, S., Dicken, C., & Kaufman, P. (2012). *Access to Affordable and Nutritious Food: Updated Estimates of Distance to Supermarkets Using 2010 Data.* Economic Research Service, United States Department of Agriculture.

Walker, R., Keane, C., & Burke, J. (2010). Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place*, *16*, 876-884.

Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2011). *Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed.* Boca Raton, FL: Chapman & Hall/CRC.

Widener, M. J., Farber, S., Neutens, T., & Horner, M. (2015). Spatiotemporal accessibility to supermarkets using public transit: an interaction potential approach in Cincinnati, Ohio. *Journal of Transport Geography, 42*, 72-83. doi:http://dx.doi.org/10.1016/j.jtrangeo.2014.11.004

Widener, M. J., Farber, S., Neutens, T., & Horner, M. W. (2013). Using urban commuting data to calculate a spatiotemporal accessibility measure for food environment studies. *Health & Place, 21*, 1-9.

Widener, M. J., Metcalf, S. S., & Bar-Yam, Y. (2013). Agent-based modeling of policies to improve urban food access for low-income populations. *Applied Geography, 40*, 1-10. doi:http://dx.doi.org/10.1016/j.apgeog.2013.01.003

Wooldridge, J. M. (2002). *Introductory Econometrics: A Modern Approach (2nd ed.)*. South-Western College Pub.

Wrigley, N., Warm, D., & Margetts, B. (2003). Deprivation, diet, and food-retail access: findings from the Leeds 'food deserts' study. *Environment and Planning A, 35*, 151-188. doi:10.1068/a35150

APPENDIX

A-1: Ordinary Least Squares Regression for Walking

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :    NAT_WALK                 Number of Observations:         224
Mean dependent var  :      1.9109                 Number of Variables   :           4
S.D. dependent var  :      0.5860                 Degrees of Freedom    :         220
R-squared           :      0.1699
Adjusted R-squared  :      0.1585
Sum squared residual:      63.577                 F-statistic           :     15.0060
Sigma-square        :       0.289                 Prob(F-statistic)     :   6.385e-09
S.E. of regression  :       0.538                 Log likelihood        :    -176.789
Sigma-square ML     :       0.284                 Akaike info criterion :     361.579
S.E of regression ML:      0.5328                 Schwarz criterion     :     375.225

-------------------------------------------------------------------------------------
            Variable     Coefficient         Std.Error      t-Statistic     Probability
-------------------------------------------------------------------------------------
            CONSTANT       1.5835125         0.1025328       15.4439547       0.0000000
            CRIMEDEN       0.0019611         0.0003491        5.6173408       0.0000001
              INLOOP      -0.2171730         0.0827632       -2.6240291       0.0092983
           NOCOLLEGE       0.0001529         0.0000440        3.4766422       0.0006117
-------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER              6.503

TEST ON NORMALITY OF ERRORS
TEST                              DF         VALUE            PROB
Jarque-Bera                        2            0.430            0.8064

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                              DF         VALUE            PROB
Breusch-Pagan test                 3            3.322            0.3446
Koenker-Bassett test               3            3.000            0.3916

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF         VALUE            PROB
Lagrange Multiplier (lag)          1           77.417            0.0000
Robust LM (lag)                    1           11.335            0.0008
Lagrange Multiplier (error)        1           66.415            0.0000
Robust LM (error)                  1            0.333            0.5639
Lagrange Multiplier (SARMA)        2           77.749            0.0000
```

A-2: Spatial Lag Regression for Walking

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :    NAT_WALK                  Number of Observations:        224
Mean dependent var  :      1.9109                  Number of Variables   :          5
S.D. dependent var  :      0.5860                  Degrees of Freedom    :        219
Pseudo R-squared    :      0.4732
Spatial Pseudo R-squared:  0.1957
Sigma-square ML     :       0.186                  Log likelihood        :   -141.800
S.E of regression   :       0.432                  Akaike info criterion :    293.600
                                                   Schwarz criterion     :    310.658

-------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
-------------------------------------------------------------------------------------
            CONSTANT       0.4026870       0.1332553       3.0219197       0.0025118
            CRIMEDEN       0.0010323       0.0002913       3.5443325       0.0003936
              INLOOP      -0.1162702       0.0669327      -1.7371215       0.0823657
           NOCOLLEGE       0.0000906       0.0000354       2.5586620       0.0105076
          W_NAT_WALK       0.6872014       0.0593504      11.5787110       0.0000000
-------------------------------------------------------------------------------------
================================ END OF REPORT ======================================
```

A-3: Ordinary Least Squares Regression for Driving

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :   NAT_DRIVE               Number of Observations:      224
Mean dependent var  :      0.8378               Number of Variables   :        4
S.D. dependent var  :      0.5803               Degrees of Freedom    :      220
R-squared           :      0.1708
Adjusted R-squared  :      0.1594
Sum squared residual:      62.264               F-statistic           :   15.1003
Sigma-square        :       0.283               Prob(F-statistic)     : 5.692e-09
S.E. of regression  :       0.532               Log likelihood        :  -174.452
Sigma-square ML     :       0.278               Akaike info criterion :   356.904
S.E of regression ML:      0.5272               Schwarz criterion     :   370.550

-----------------------------------------------------------------------------------
       Variable     Coefficient       Std.Error      t-Statistic     Probability
-----------------------------------------------------------------------------------
       CONSTANT       0.5080266       0.1014685       5.0067417       0.0000011
       CRIMEDEN       0.0019616       0.0003455       5.6777044       0.0000000
         INLOOP      -0.2047959       0.0819041      -2.5004366       0.0131336
      NOCOLLEGE       0.0001497       0.0000435       3.4382088       0.0007002
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER         6.503

TEST ON NORMALITY OF ERRORS
TEST                            DF        VALUE            PROB
Jarque-Bera                      2          0.599          0.7413

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                            DF        VALUE            PROB
Breusch-Pagan test               3          3.343          0.3417
Koenker-Bassett test             3          2.967          0.3967

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                          MI/DF       VALUE            PROB
Lagrange Multiplier (lag)        1         83.289          0.0000
Robust LM (lag)                  1         10.649          0.0011
Lagrange Multiplier (error)      1         72.764          0.0000
Robust LM (error)                1          0.123          0.7256
Lagrange Multiplier (SARMA)      2         83.412          0.0000
```

A-4: Spatial Lag Regression for Driving

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :   NAT_DRIVE               Number of Observations:       224
Mean dependent var  :      0.8378               Number of Variables   :         5
S.D. dependent var  :      0.5803               Degrees of Freedom    :       219
Pseudo R-squared    :      0.4898
Spatial Pseudo R-squared:  0.1935
Sigma-square ML     :       0.177               Log likelihood        :    -136.932
S.E of regression   :       0.421               Akaike info criterion :     283.865
                                                Schwarz criterion     :     300.923

-----------------------------------------------------------------------------------
            Variable    Coefficient       Std.Error     z-Statistic    Probability
-----------------------------------------------------------------------------------
            CONSTANT      0.0581832       0.0894870       0.6501858      0.5155722
            CRIMEDEN      0.0010218       0.0002846       3.5905707      0.0003300
              INLOOP     -0.1109909       0.0652537      -1.7009122      0.0889595
           NOCOLLEGE      0.0000877       0.0000345       2.5386290      0.0111288
         W_NAT_DRIVE      0.7017679       0.0575829      12.1870918      0.0000000
-----------------------------------------------------------------------------------
================================= END OF REPORT ===================================
```

A-5: Ordinary Least Squares Regression for Transit

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :   NAT_TRANS                Number of Observations:        224
Mean dependent var  :     2.2027                 Number of Variables   :          4
S.D. dependent var  :     0.4301                 Degrees of Freedom    :        220
R-squared           :     0.2088
Adjusted R-squared  :     0.1981
Sum squared residual:     32.634                 F-statistic           :    19.3582
Sigma-square        :      0.148                 Prob(F-statistic)     :  3.552e-11
S.E. of regression  :      0.385                 Log likelihood        :   -102.098
Sigma-square ML     :      0.146                 Akaike info criterion :    212.195
S.E of regression ML:     0.3817                 Schwarz criterion     :    225.842

------------------------------------------------------------------------------------
            Variable    Coefficient      Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      2.2124535      0.0734598      30.1178903       0.0000000
            CRIMEDEN      0.0012121      0.0002501       4.8461214       0.0000024
              INLOOP     -0.4002352      0.0592958      -6.7498085       0.0000000
           NOCOLLEGE      0.0000781      0.0000315       2.4785458       0.0139437
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           6.503

TEST ON NORMALITY OF ERRORS
TEST                              DF        VALUE           PROB
Jarque-Bera                        2         2.373           0.3053

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                              DF        VALUE           PROB
Breusch-Pagan test                 3        10.796           0.0129
Koenker-Bassett test               3        11.287           0.0103

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                            MI/DF       VALUE           PROB
Lagrange Multiplier (lag)          1        94.440           0.0000
Robust LM (lag)                    1        15.165           0.0001
Lagrange Multiplier (error)        1        79.888           0.0000
Robust LM (error)                  1         0.614           0.4335
Lagrange Multiplier (SARMA)        2        95.053           0.0000
```

A-6: Spatial Lag Regression for Transit

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            : CTproj1.dbf
Weights matrix      :File: CTprojshp1.gal
Dependent Variable  :   NAT_TRANS                Number of Observations:       224
Mean dependent var  :      2.2027                Number of Variables   :         5
S.D. dependent var  :      0.4301                Degrees of Freedom    :       219
Pseudo R-squared    :      0.5164
Spatial Pseudo R-squared:  0.2513
Sigma-square ML     :       0.092                Log likelihood        :   -62.488
S.E of regression   :       0.303                Akaike info criterion :   134.976
                                                 Schwarz criterion     :   152.034

------------------------------------------------------------------------------------
            Variable    Coefficient      Std.Error    z-Statistic    Probability
------------------------------------------------------------------------------------
            CONSTANT      0.6273727      0.1437587      4.3640669      0.0000128
            CRIMEDEN      0.0006601      0.0002019      3.2687442      0.0010803
              INLOOP     -0.1591520      0.0491671     -3.2369642      0.0012081
           NOCOLLEGE      0.0000478      0.0000248      1.9269088      0.0539910
         W_NAT_TRANS      0.6936040      0.0587099     11.8140882      0.0000000
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```